NTNU

Kunnskap for en bedre verden

Department of Computer Science

TDT4259 - Applied Data Science

Examining Regulatory Impact on Hosting Practices in New York - Airbnb

Authors:

Name	**Student ID**	**Email**
Malin Døskeland	527204	malinsd@stud.ntnu.no
Camilla Kopperud	527217	camillkk@stud.ntnu.no
Olav Utle Kolltveit	506682	olavuk@stud.ntnu.no
Magnus Johansen Pierce	527796	magnuspi@stud.ntnu.no
Mikkel H. Sandhaug	527759	mikkelhs@stud.ntnu.no
Heidi Therese Wiest	507396	heiditwi@stud.ntnu.no

2023-11-27

Table of Contents

List of Figures

List of Tables

Part 1

Introduction and problem definition

This report is the final result of a group assignment for the course TDT4259 at the Norwegian University of Science and Technology (NTNU). Our group has elected to focus on Airbnb's operations in New York City (NYC), using data science to assist the company in identifying trends on their platform and adapting to the new regulatory developments and market conditions. Drawing upon publicly available datasets of Airbnb listings and the geographic locations of public subway systems, we aim to provide a data-driven perspective that can support strategic business decisions. Our approach is based on design thinking and the CRISP-DM framework, ensuring a methodical progression from understanding business needs to delivering data-informed recommendations.

1.1 About Airbnb

Airbnb is a popular online marketplace that connects individuals looking to rent out their homes with people seeking accommodations in the area. The platform offers a large variety of properties, including single rooms, suites, apartments, houses, and houseboats. Hosts create the listings on the platform by providing detailed information about the property, including photos, pricing and availability. Renters browse the listings and book their desired property. Airbnb facilitates the payment through the platform and takes a commission from both parties. Airbnb also allows for reviews of both hosts and renters, adding a layer of trust to the platform (Airbnb, 2023b).

1.2 Context and domain

For several years there has been an ongoing feud between cities and home-sharing companies. NYC has argued that short-term rentals - classified in this report as rentals with a duration of 10 days or less - via platforms like Airbnb contribute to escalating rents and the current housing shortage. On the other hand, Airbnb and similar platforms claim to strengthen the city's tourism industry. For several years already, there have been regulations for short-term rental in NYC. These rules stated that hosts are required to be present during stays of less than 30 days and guests must have access to the entire facility. Additionally, there was a guest cap of two. However, these regulations did not prove to be effective as hosts continued to rent out their properties illegally. Airbnb has been accused, by city officials, of not enforcing these regulations strictly enough and deriving as much as half of its revenues from unlawful renting activities. However, this claim has been contested by Airbnb (Zaveri, 2023).

Due to the issues related to the previous restrictions, NYC started enforcing new regulations this September, which significantly restrict the rental opportunities for hosts. For hosts to continue their short-term rental activities, they must register with the city for short-term rental authorization (Chan, 2023). Airbnb are required to confirm their registration status before collecting any fees from hosts. If not, they can expect significant fines. This will lead to a substantial amount of short-term rental listings being forcibly removed or transitioned to having a minimum night stay of 30 days. This will put many displaced hosts in the position of having to adjust their rental prices and practices to an unfamiliar rental segment. Airbnb describes the regulations as a “de facto ban” (Inside Airbnb, 2023a) and are facing severe uncertainties related to their future platform activity and financial performance in NYC in the aftermath of the regulations.

1.3 Motivation

The primary revenue source of Airbnb is commissions, and their earnings are therefore likely to be significantly impacted by the removal of short-term listings and the potential suboptimal rental practices of new inexperienced medium-term hosts. This leaves Airbnb in a state of revenue uncertainty. Due to this, Airbnb should conduct an analysis of the regulations' impact and create a strategic plan to effectively navigate in the new rental environment. Fortunately, Airbnb has a dataset of their listings, which includes in-depth information on NYC property listings, user profiles, and reviews. A thorough data science analysis utilizing this dataset can be key factor for Airbnb to navigate the current uncertainties and ensure a bright future.

These regulatory challenges are not an isolated incident. Similar regular scrutiny is seen and expected in several cities worldwide (O'Sullivan, 2023). Therefore, the methodology and insights from this NYC case also has significant relevance for other cities globally. The knowledge acquired in this project can therefore be used to create a globally applicable strategy for regulatory compliance and market adaptation.

1.4 Problem description

When selecting the problem description, it is important to align it closely with the organization, its current environment, and the dataset at disposal. For maximized effectiveness, the problem should address one of the organization's most pressing challenges. In the case of Airbnb, we have identified the following problem:

> **Problem description**: How can data science be used to understand the effect of recent changing market dynamics in the NYC rental market, and how can Airbnb utilize this insight to effectively handle the changes?

Consequently, this report aims to analyse the changing market dynamics in NYC and provide actionable insights to help Airbnb adapt appropriately and effectively.

There were two primary factors in our decision to choose this project description. Most importantly, the regulatory changes directly affect Airbnb and it is one of their most pressing challenges today. In addition, Airbnb has a comprehensive dataset on property listings in NYC, which gives us the necessary tools to address the problem effectively.

1.5 Team roles and responsibilities

The team possesses a diverse set of skills and knowledge relevant for the project which they have developed through studying Industrial Economics and Technology Management and Computer Science at NTNU and in work placements. The combined experience of all the team members include data analysis, machine learning, business and finance. The motivation for the team is to gain a comprehensive understanding of data science projects and how they align with real-world business cases. To ensure the success of the project the team will focus on aligning the project in accordance with the problem description and motivation described in Section 1.3 and 1.4 respectively, in addition to prioritize sharing of knowledge amongst the members.

To ensure a successful project the team prioritised developing a shared mental model of the project early on. To create the shared mental model all members were involved with researching both Airbnb and the data. Additionally, the team distributed responsibilities to ensure all parts of the project was done in a satisfactory manner and to avoid putting all responsibilities on a single individual. The team members were not restricted by their responsibilities and contributed to other aspects of the project in addition to their area of responsibility. The responsibilities, roles and team members are presented in Table 1.

Table 1: Overview of the team members and their corresponding information and responsibilities

Member	**Background**	**Responsibilities**
Camilla	Proficient in machine learning, visualization, and coding	Business understanding, project design, and video creation
Heidi	Organizational management and ML expertise from internships	Preprocessing and analyzing data, creating visuals
Malin	Interested in how data converts to business improvements	Data cleaning, detailed analysis, and visualization
Magnus	Little experience with machine learning, experienced in group work	Understanding business needs, model creation
Mikkel	Experienced in machine learning for time series projects	Business insight, project planning, and video production
Olav	Aims to deepen his data science knowledge, proficient in machine learning already	Preparing data, analytical research, and analyzing results

Part 2

Background

This section of the report will provide information on the objectives of the project. The choice of the data science project management and data strategy will also be presented and how they were implemented.

2.1 Objectives

After formulating the problem statement in Section 1.4, the team began brainstorming possible objectives. Both the understanding of the business and the data was explored. The understanding of the business contributed to finding relevant objectives to Airbnb and to evaluate their value for the company. The initial data exploration was performed to give insight into the feasibility of each of the possible objectives that were identified by the team. The objectives identified by the team are presented in Table 2, together with whether or not they were ultimately chosen after an iterative process.

Table 2: Objectives

ID	Objective	Chosen
1	Confirm regulatory impact.	Yes
2	Identify the regulatory impact on affected customer segments.	Yes
3	Assist new medium-term hosts through specific pricing proposals.	Yes
4	Provide insights into the short-term costs of regulations assuming they subside.	No
5	Develop a general pricing model.	No
6	Assess the added value of amenities on pricing.	No
7	Determine the expected cost of a tenant during a specific stay.	No

The objectives will now be further explained. Their business value and feasibility will be evaluated and further presented in Figure 1.

2.1.1 Confirm regulatory impact.

Assess if the anticipated increase in minimum nights for short-term rentals occurs post-regulation. To assess this, data on listings both from before and after the regulations need to be examined. This analysis is highly feasible since the team possesses the relevant listing data. The objective also has a high business value because it confirms whether or not the regulations have actually forced hosts to change their status to medium-term hosts, which could further indicate to Airbnb that their business may be affected.

2.1.2 Identify the regulatory impact on affected customer segments.

To further inform Airbnb about the consequences of the new regulations, the team identified the possibility of finding information about which customer segments are the most affected by the regulations and how significantly they are affected. This holds a high business value for Airbnb, as this allows the company to concentrate on counteracting the regulations' effect on the identified customer segments to retain as many of their customers as possible. The data needed would be the listings which are affected. The team possesses the necessary data, making the feasibility high.

2.1.3 Assist new medium-term hosts through specific pricing proposals.

The aim is to create a model that not only provides precise pricing recommendations for hosts shifting from short-term to medium-term rentals due to regulatory changes, but which also explains the rationale behind these prices. Recognizing the uniqueness of this situation, the team proposed a specialized pricing model. Utilizing data on pricing and availability, a feasible model can be constructed for these medium-term hosts. The perception of the model's business value has evolved from low to high as detailed in Figure 1. Initially, we thought a specialized pricing model for only a subset of hosts would give limited business value, especially since we believed the new regulations would have only a short-term impact on the rental market. However, after reassessing the permanence of the regulatory changes, as discussed in Section 4.4.5, and further quantifying the potential revenue losses during the initial analysis, we believe this objective to have considerably high business value.

2.1.4 Provide insights into the short-term costs of regulations assuming they subside.

This objective would provide insight into how long Airbnb should expect a revenue decrease in New York, assuming the regulations eventually subside and the old balance is restored. This will leverage insights from regulatory experiences in Jersey City and North Carolina to estimate the duration of the effects for New York. Hence, data will be needed for Jersey City, North Carolina and New York. During the first evaluation phase, the team recognized that our initial assumption of the regulatory setbacks being temporary was faulty, which drastically decreased the potential business value for Airbnb, as shown in Figure 1. For more details, read Section 4.4.1. The feasibility always remained high, as the required data for Jersey City and North Carolina is available through the same source as for New York (Inside Airbnb, 2023b).

2.1.5 Develop a general or area-specific pricing model.

This objective would consist of making a general pricing model which could give all landlords around the globe price suggestions for their rentals. The objective is one of the team's first ideas, as it would be feasible given the data and would be valuable for the business to increase Airbnb's robustness. However, after the initial data exploration the team realized that a general pricing model would be infeasible given the hundreds of datasets for different cities which would need to be merged. The alternative would be to create a specialized pricing model for a limited area only. We believe this alternative approach is more feasible, but in reality is just an inferior version to the model in Objective 3, which directly targets hosts affected by the regulations. 1 therefore shows that the area-specific pricing model is assigned low business value.

2.1.6 Assess the added value of amenities on pricing.

This represents a different goal than price prediction, as we instead would be interested in understanding how hosts can increase their prices by implementing simple changes through buying the most important amenities they are lacking. This objective could increase Airbnb's earnings, making them more robust against the potential dangers of the new regulations. However, the business value was regarded as quite low, as we believe amenities are not impactful enough on the price to create any significant change. This was later confirmed by the feature importance from the pricing model in Figure 14. On the other hand, the feasibility was set to quite high as the team had data on the amenities included in a specific listing.

2.1.7 Determine the expected cost of a tenant during a specific stay.

To enhance tenant satisfaction using Airbnb and increase the customer loyalty amongst tenants to Airbnb, the objective was to ascertain the expected cost of a stay for a tenant. The only data the team possesed was data related to the host perspective, which made the feasibility of this objective rather low. The business value is also comparatively low, as it does not substantially counteract Airbnb's immediate regulatory issues or contribute to the overall robustness of the platform in terms of rapid revenue improvements.

2.1.8 Choice of objectives

When choosing the final objectives for the project, the team chose the objectives with both high business value and high feasibility. In Figure 1 these are presented in the green square, namely the objectives with ID 1, 2 and 3 from Table 2.

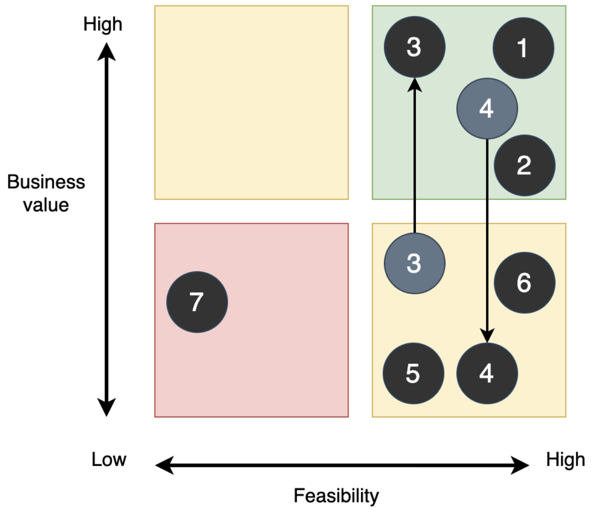


Figure 1: Feasibility Matrix.

2.2 Data science project management and data strategy

In choosing how to manage and strategize this project the team wanted to explore widely used methods, preferably relevant to industry. The team also wanted to have the opportunity to adjust the project along the way as a consequence of new findings, and we therefore aimed for agile methods which allows for adjustments and several iterations if necessary, further described in Section 4.4. To manage the data science project the team chose Cross Industry Standard Process for Data Mining (CRISP-DM) as it is most commonly used (Saltz, 2022) and can be implemented as an agile process with several iterations, since you can go back to earlier steps, see Figure 2. Since Design Thinking is an iterative process (Dam & Siang, 2022) it fits well with the team's aim for an agile process, and adds value to the data strategy as it makes sure the team has the user in focus. This will add to the likelihood of the success of the project as to giving the customer, Airbnb, what they need and want.

2.2.1 CRISP-DM

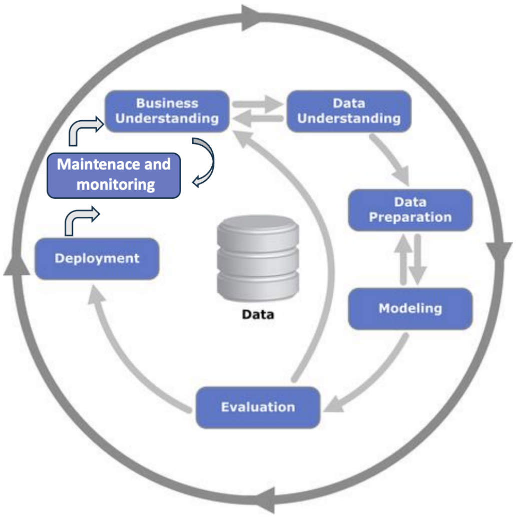


Figure 2: CRISP-DM.

2.2.1.1 Business understanding

To initiate the project, the team sought to understand Airbnb in New York, as the target business for this project. This is commenced with the question: “What does the business need?” (Hotz, 2023a). To answer this query, the team researched relevant literature both as to what Airbnb is and whether they have any problems or opportunities that could be leveraged. The team incorporated Design Thinking, as described in Section 2.2.2, to complement this process.

2.2.1.2 Data understanding

To explore potential value creation opportunities, the team examined available data. First the team collected data from Inside airbnb (Inside Airbnb, 2023b) and data on the geographical locations of subways in New York from the NYC Open Data (New York Open Data, 2023).The team described and explored the data to gain insights, aiming to identify relationships and highlight interesting aspects. They then verified the data quality. Additionally, literature related to data science projects and rental units was explored.

2.2.1.3 Data preparation

Following data collection, the team prepared the gathered data. First the team selected the relevant data from Inside Airbnb(Inside Airbnb, 2023b) and NYC Open Data (New York Open Data, 2023). Thereafter the data was cleaned, constucted, integrated and formatted to make it prepared for modelling.

2.2.1.4 Modeling

The team started with the question, ”What modeling techniques should we apply?” (Hotz, 2023a). To answer this the team looked at the project’s objectives and relevant literature. They explored several models by first building them and then assessing them as to how they fitted for their purpose. This was done to select the most suitable one.

2.2.1.5 Evaluation

Models were evaluated based on business success criteria, with the team determining whether they were deemed adequate for deployment or if another iteration in CRISP-DM was necessary to better align with the business case. The process was reviewed and the next step was determined. In this project there were a need for several iterations,this is described in greater details in Section 4.4.

2.2.1.6 Deployment

In this phase, the team planned the deployment processes tailored to the project’s objectives. Recommendations and an implementation plan was designed. These were designed to enable Airbnb to incorporate the value generated by the project.

2.2.1.7 Maintenance and monitoring

In this phase the team recommend to monitor the deployment by using several metrics, and have them set off alarms when the measure of a metric deviates significantly from the expectation. To maintain the business value, maintenance is recommended.

2.2.2 Design Thinking

In parallel with the CRISP-DM lifecycle, the team employed Design Thinking, as depicted in Figure 3, to emphasize placing stakeholders and users at the center throughout the data science project. Because design thinking ”seeks to understand your users” (Hotz, 2023a), the business understanding of Airbnb was significantly enhanced as the team empathized the project's users—both Airbnb as the direct customer and, in turn, their users, the landlords, and the tenants. Later in the CRISP-DM lifecycle, design thinking contributed to ensure that wanted approaches, developed by defining a problem and coming up with ideas for a solution, and then prototyped and tested, would satisfy the team's customer, AirBnb, and their customers. The approach was utilized iteratively to find objectives, look at their feasibility, value and prototype through plots. An example is how the findings related to objective 2, Section 2.1.4, illuminated by prototyping and testing, revealed that the changes were more permanent than expected. This prompted the team to reassess the business understanding in CRISP-DM, and redefined the base of the ideas for the Design thinking. Subsequently, the team ideated again and enhanced the value of objective 5.

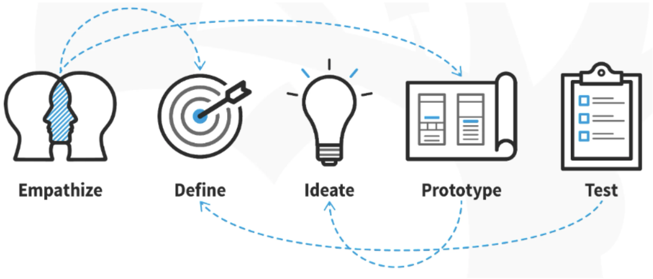


Figure 3: Design Thinking.

2.3 Relevant examples

While researching relevant literature, the team aimed to investigate models previously employed in predicting rental prices, aligning with the project objectives. One instance involved the application of regression models for predictions of rental prices for apartments in Brazil (Shah et al., 2021). Another case entailed a comparison between a regression model and a neural network. The findings indicated that the regression model showed greater accuracy for smaller sample sets, and the expectation of the neural network being better suited for bigger sample set (Seya & Shiroi, 2022). Consequently, both regression models and neural networks are techniques explored in the project.

Part 3

Methods

3.1 Data

Description

The main datasets used for this project are the datasets of all listings in NYC from November 2022 to October 2023 sourced from Inside Airbnb (2023b), which is a website that scrapes data from the Airbnb platform. Each month, the website publishes an updated dataset with the current listings for multiple cities around the world. The chosen datasets consist of more than 38 000 listings each, in which the columns of the datasets can be divided into two primary categories: host-associated and listing-specific columns. There are 14 host-associated features where the majority of them are numerical. These features and their explanations can be found in Appendix A. There are 32 listing-specific features where the columns are a mix of numerical, categorical and plain text attributes. These features and their explanations are given in Appendix B.

In order to collect more relevant information which can be used to predict pricing, another dataset consisting of subway stations in NYC is also used. This is a public dataset provided by the New York Open Data (2023). The dataset consists of the location, stop name and line information for each subway and railway station in the same area. In this project we are only interested in the coordinates of the stations.

Validity

Some of the columns from the main datasets are cleaned by Inside Airbnb as indicated by the 'cleansed' tag. However, we will see that there are still a great number of columns with missing values. Since each column contains an url linking to the listing online, we were also able to confirm the validity of the columns' values. For every listing we checked manually, the provided dataset information was identical with that on the Airbnb website.

One important consideration is whether the data provided by Inside Airbnb is presented in a biased manner in order to encourage increased regulation pressure towards Airbnb. The people behind Inside Airbnb appear to have a clear agenda based on the statements made on the website, such as how Airbnb are destroying the long-term rental market (Inside Airbnb, 2023a). Therefore, Inside Airbnb may choose to only make certain datasets publicly available in order to justify their own beliefs.

3.2 Methods and Tools

Throughout the whole project, a variety of tools are used. These include Python and several libraries such as Pandas, Matplotlib and Seaborn. In this section we describe how we utilize these tools in our project.

Python

Python is a popular, high-level programming language equipped with numerous of libraries for data analysis and visualization (Pafitis, 2022). The language promotes code readability and is easy to understand. Python was selected for this project due to the comprehensive analysis tools it offers. Our extensive experience with Python, gained from various other projects, made this language well-suited for our task.

Jupyter Notebook

Jupyter Notebook is an open-source interface for interactive Python development that allows users to create and share documents. It is particularly favored in data science for its ability to combine code execution with rich text elements (Jupyter, 2023). The project's processes were carried out entirely in Jupyter Notebooks, chosen for its speed and simplicity.

Pandas

Pandas stands as one of the most utilized libraries in data science because of its data manipulation and data analysis tools (NVIDIA, 2023). Together with Jupyter, Pandas creates a robust environment for data exploration, processing and manipulation. We chose to use Pandas because of its user-friendly nature and its impressive speed, which is particular crucial when dealing with large datasets.

Matplotlib and Seaborn

Matplotlib and Seaborn are two Python libraries frequently used for data visualization. Both libraries offer a wide range of options to create complex plots and figures. Seborn is built on top of Matplotlib and has the advantage of handling massive amounts of data with ease (Atanda, 2023). Both of these libraries were chosen for this project because they are easy to use and provide tools for making beautiful visualizations.

Machine Learning

Machine Learning models employ algorithms and statistical methods to identify patterns within data and make decisions without explicit programming. This approach is particularly beneficial for our project since we have a large dataset from Airbnb that is well-suited for machine learning analysis. There are various machine learning models to choose from, each offering distinct features and potential uses. For our purposes, we will be utilizing a Random Forest Regressor.

Random Forest Regressor A Random Forest Regressor uses many decision trees to compute an average prediction This enhances both prediction accuracy and robustness by combining multiple weak decision trees into a strong model. In contrast to many other models, it offers a higher degree of transparency, providing the ability to evaluate feature importance effectively.

Despite these benefits, Random Forests can be computationally demanding and are less transparent than a single decision tree, with complexity increasing with the number of trees.

3.3 Preprocessing

The first step of the preprocessing is to get familiar with the listing datasets and to select the appropriate features for the task. The preprocessing is conducted with all the objectives in mind, but places an emphasis on the cleaning done to ensure that the dataset was ready for the machine learning model. All the listing datasets consist of 75 columns with a unique id as identifier. In this section we explain how we cleaned the listing dataset from October 2023. The preprocessing is done in the same way for all other datasets. In the first data cleaning phase columns were removed according to the following rules:

1. Columns containing irrelevant information. In this category fall features such as URLs, identical scraping information for each listing, and non-unique identifiers of the hosts.

2. Columns containing similar information in text-format as other categorical and numeric features. This involves the features named 'neighborhood_overview', 'host_about', and 'host_verifications'.

3. Columns containing only missing values. This includes the bathroom column and the license column.

4. Columns containing the same information as the cleaned columns provided by Inside Airbnb. Here we removed the neighbourhood and the host listing count columns.

5. Columns offering the same value as other columns. The features 'first_review' and 'last_review' indicates host experience that mirror the same value as the 'host_since' column. Since the first and last review columns in addition contain many missing values, these are removed from the dataset.

The remaining columns can be divided into two groups based on the information they give. These two groups are features related to the specific listing and features associated with the belonging host. The name and description of the remaining features can be found in Appendices A and B. For both of the groups, the columns with missing values are illustrated in Figure 4.

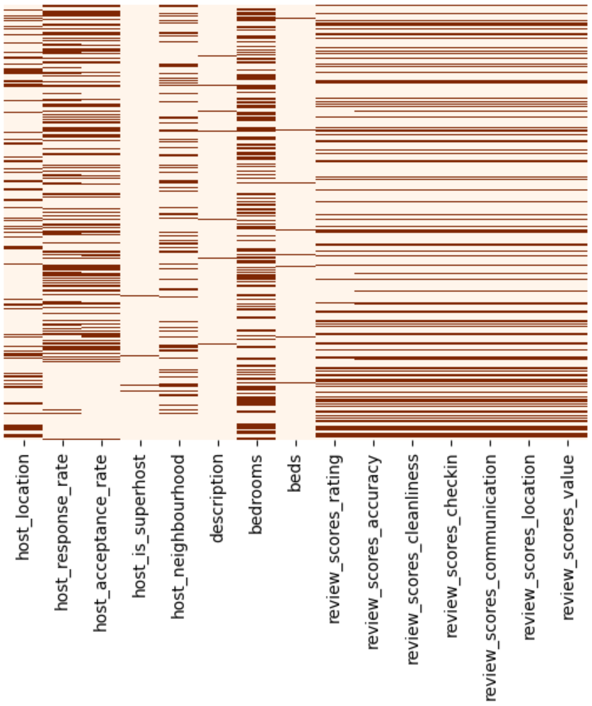


Figure 4: A visualization of missing values for all features. The missing values are coloured in red.

Host-associated features

As seen from Figure 4, only five of the host-associated features contain missing values. The following part of this section describes the modifications made to these features.

Host Location and Neighbourhood As shown by Figure 4, the 'host_location' and 'host_neighbourhood' columns have many missing values. Since many of the rows lack data in only one of these columns, it seems like there is a potential to impute a portion of the 'host_location' missing values using the other column. On closer analysis, it became apparent that neighbourhood identifiers are not unique to individual cities, which renders this approach moot. However, the final methodology used was to combine the two columns into one based on the following rules:

1. If both columns have a value, then the result will be a tuple of the neighbourhood and its general location.
2. If only the location is given, the location should be the value of the new column.
3. If only the neighbourhood is given, there is not enough information to pinpoint a host location and the value is therefore set to 'Ambiguous location'.
4. If both columns have missing values the value will be 'Not disclosed'.

Superhost To address the missing values in the 'superhost' column, a sample of records were manually checked out. It was consistently observed that none of the hosts in these instances qualified as superhosts. Consequently, the missing entries were filled with zeros.

Host response and acceptance rate The host response rate is explored further in the second iteration of the project. Initially, the feature was removed due to the high occurrence of missing values as seen from Figure 4. However, multiple members voiced the opinion that the feature may impact the pricing of a listing, since a host who offers superior communication with guests may charge a premium for this. When comparing the missing values of the host response rate to the missing values of the host acceptance rate, it became clear that many of the host response rates for listings are missing completely at random with no proper explanation. Multiple listings have a value greater than zero for the acceptance rate, implying they have responded positively to a booking, yet have a missing value for the response rate. Hence, any attempt to impute missing values appears futile, and so the feature is once again dropped.

Listings-specific features

Figure 4 shows that ten of the listings-specific columns contain missing values. The reminder of this section specify the preprocessing applied to this category of features.

Bedrooms and beds Many of the listings rented out have missing values in the bedrooms column, when the room type states that a single room is rented out. Substituting the missing values with 1 effectively eliminated most of them from the column. In order to fill the remaining missing values, we first assumed that the number of people the listing can host is a decent proxy for the number of bedrooms. The rest of the missing values are therefore filled by the mean value of the bedrooms column grouped by the values of the accommodates feature. The same approach is used for the 'beds' column.

Rating features For the initial attempt at completing Objective 3, all features relating to various ratings of a listing were included in the price prediction model. The hypothesis was that each review feature described a distinct aspect of a stay (communication, cleanliness, etc...) and so the belief was that only using the overall rating would not perform as well for the pricing model. Missing values were imputed by understanding that each occurrence of a missing value for any rating feature was associated with a value of zero for the 'number_of_reviews' feature. Hence, missing values for the rating features clearly indicated a lack of reviews. The missing values were replaced with the median over all observable ratings for a given rating feature. If we instead had used zero or "-1" to impute these values, then it would have likely skewed the ratings heavily seeing as the median and mean for all ratings features were in the interval 4.6-4.8 for the month of October.

For the final attempt at the price prediction model, the ratings features were revisited by exploring the correlation matrix seen in Figure 5. A correlation matrix helps to explain the co-movements between selected variables, and a strong correlation may typically be defined as 0.8 or greater. The figure made it apparent that the prediction models may be suffering from multicollinearity, as nearly all the ratings features have correlations of 0.7 or higher. We assume that the strong correlations are caused by the fact that many guests simply do not bother to give detailed reviews after a stay, and instead choose to give the same score for all aspects of a listing. Based on these insights, we choose to remove all rating features except the overall rating when predicting prices for Objective 3.

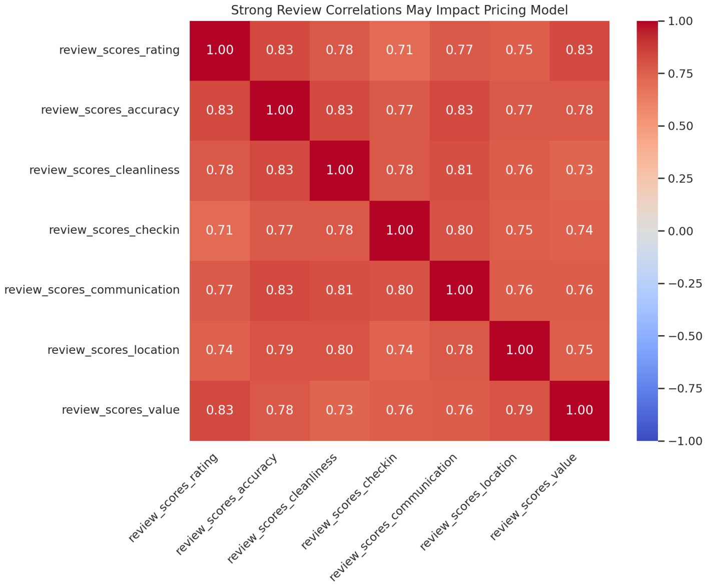


Figure 5: A matrix showing the correlation between the different rating features.

Property type and amenities Originally, the property type was divided into 81 unique string values. In many machine learning models, categorical features are automatically one-hot encoded, meaning that each unique value of the property type would get an additional column. This may lead to a phenomenon known as the "curse of dimensionality" where the feature space becomes so large that the model requires an excessive amount of data to learn effectively. To mitigate this effect, the unique values are sorted in ascending order according to the average price for each category. The original column is then replaced by this ordering. A similar one-hot encoding process is also applied to the amenities feature, with an 'amenities_count' feature also being created in case of dimensionality issues.

Price The price column in the dataset exhibits a left-skewed distribution, which can pose challenges for machine learning algorithms that assume normality of the input data. To address this skewness, a logarithmic transformation was applied to the price column. The result can be viewed in Figure 6.

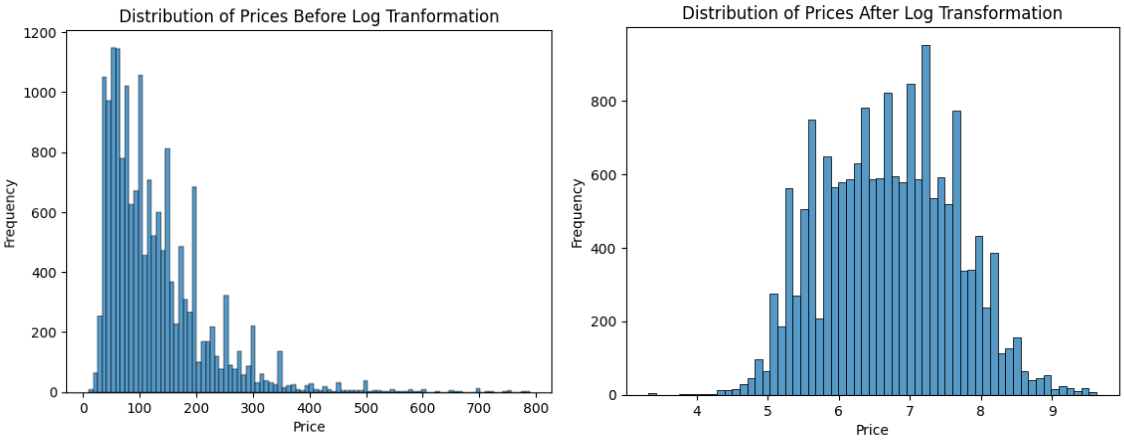


Figure 6: The distribution of the price column before and after the log trasformation.

Feature extraction

As a part of final iteration of the modeling for Objective 3, new potential features were explored. Based on the past experiences of group members, it was determined that a feature relating to ease of subway or railway access may impact the pricing of a listing. The reasoning behind this is that many of NYC's most popular attractions are situated in Manhattan, such as the Empire State Building or the Rockefeller center. A tenant in Brooklyn or Queens may value a close proximity to subways or railways in order to more easily reach points of interest, which could positively impact the price of the listing. The primary dataset does not contain any specific information related to public transportation, which introduced the need for an additional data source. Using the dataset containing information about subway and railway stations, the distance to the closest station for each listing is computed using the coordinates given in both the listings dataset and the subway dataset.

Part 4

Analysis

This section shows results and insights obtained from the data analysis, following the methodologies outlined in Section 3. Objective outlined in Section 2, will be chronologically addressed, detailing our approach and discoveries. Afterwards we will reflect on the findings and discuss how our project evolved over time. The insights garnered form the basis for the recommendations proposed in Section 5.

4.1 Objective 1: Confirming the Effect of NYC Regulations

As stated prior, the new regulations will force previous short-term hosts to adopt a minimum night stay of 30 for all their listings which are not approved by the NYC state. Upon plotting the proportion of listings according to the minimum number of nights, as seen in Figure 7, a noticeable transition is observed from hosts renting out accommodations for less than 30 nights to a minimum of 30 nights, between August and October 2023. During the months prior to August 2023, the portions between the number of minimum nights below and equal to 30 nights have been stable. This correlates with our understanding of the impact of the September regulations. Figure 7 also shows the stable trend in listings with a minimum night stay above 30 for the whole period. These findings show the significant impact the regulations have had on the data.

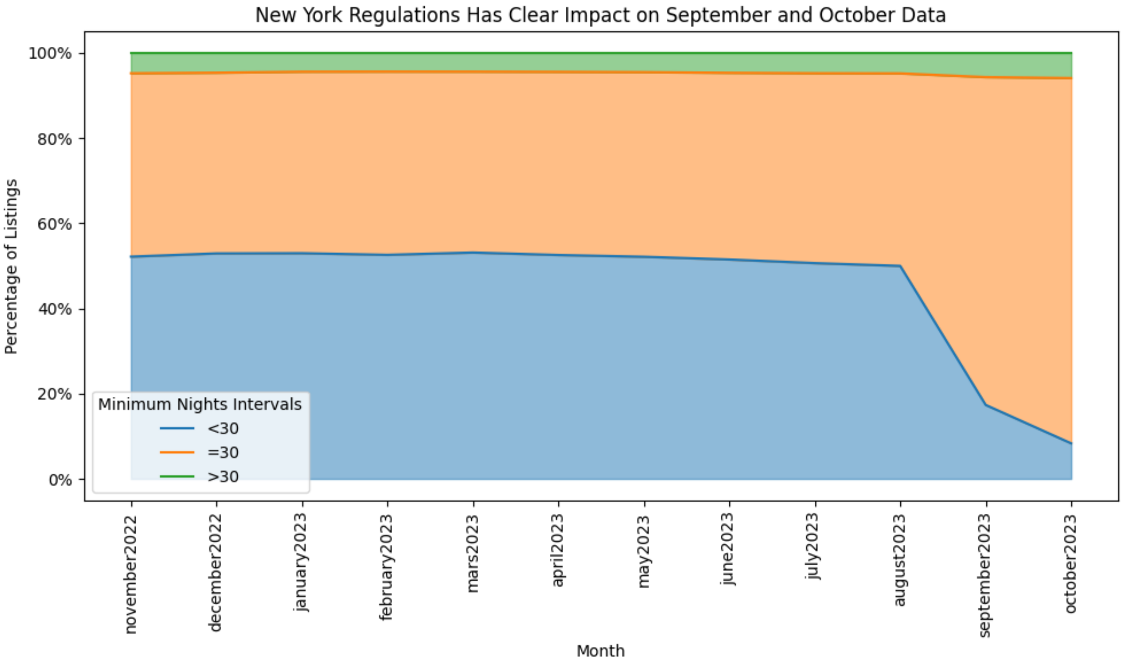


Figure 7: A visualization of the portion of listings with minimum nights less than, equal to and more than 30 days for the last year.

4.2 Objective 2: Identify the regulatory impact on affected customer segments

For the second objective, the first goal is to identify the hosts which are most likely to feel the effects of the new regulations. We find that many affected hosts appear to be regular people with only one listing, who may simply wish to share their homes with others. The second goal is to quantify the effects of the regulations, which is done by initially assuming that affected hosts are currently struggling to price their listings in the 30 minimum nights market, hereafter referred to as the 30-day market. Our findings suggest that potentially millions of dollars in revenues will be lost for impacted hosts, unless immediate steps are taken to correct their prices.

4.2.1 Identifying Host Segments Impacted by Regulations

In order to identify exactly which hosts have been affected by the regulations, previous short-term hosts are first isolated from the October dataset. This is done by determining the IDs of all hosts which have moved from the short-term to the 30-day rental market from August. We find that over 12’000 hosts seem to have been affected by the recent regulations. In other words, there are thousands of displaced hosts who are now forced to compete in a medium-term market they are inexperienced with.

When analyzing the segmentation of the affected hosts, roughly 60% only have a single listing on Airbnb as seen from Figure 8. This indicates that affected hosts are generally regular people who are renting out their rooms or apartment. Conversely, very few of the displaced hosts seem to be ”professional” hosts who seek to buy up multiple properties purely for rental purposes. The main justification for the regulations seems to be that it makes it less enticing for large rental companies to buy property in NYC for short-term hosting (Zaveri, 2023). However, our insights from Figure 8 indicate that many regular hosts are affected instead, which raises some doubt as to whether the new regulations are impacting the intended target.

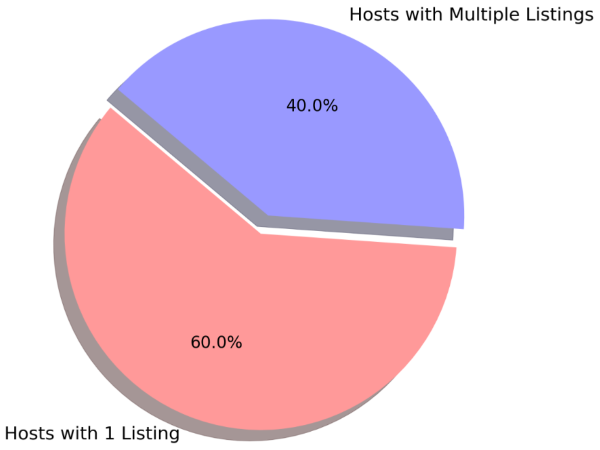


Figure 8: The percentages of displaced short-term hosts with only 1 listing as opposed to multiple listings.

Figure 9 shows that the most frequent listing category for affected hosts is ”entire rental unit”. However, when scanning the rest of the chart, it is evident that the majority of affected hosts are renting out a private room. This further indicates that the affected hosts are regular people, many of whom simply wish to rent out a part of their home. This finding also means that many of the affected hosts may struggle to return to the short-term rental market, as the new regulations clearly state that guests must have access to the entire home for short-term stays (Zaveri, 2023).

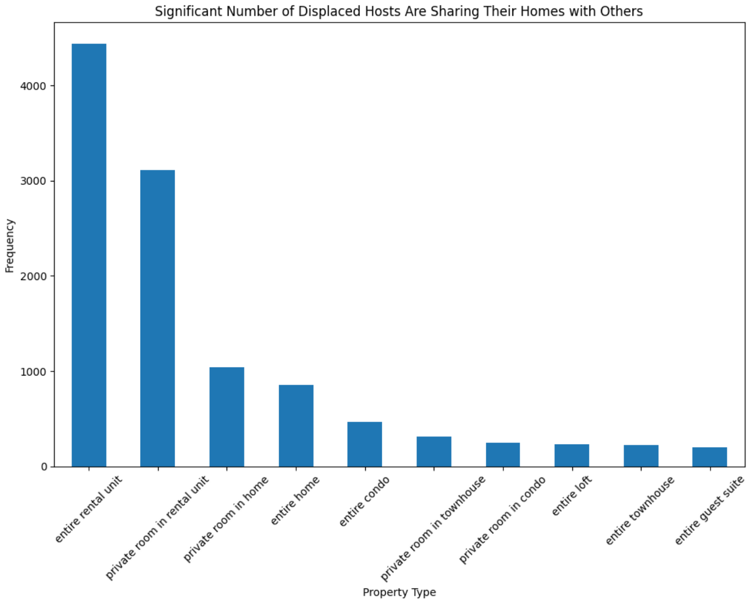


Figure 9: The occurrences of the most frequent property types among displaced hosts

4.2.2 Prices and Booking Rates in the Short-Term and Medium-Term Markets

Now that we have determined which host segments that are affected by the regulations, the next step is to understand exactly how their experience with Airbnb may change. As part of the ”Empathize” and ”Design” stages of the Design Thinking process, we wanted to understand the most important drivers for Airbnb and their hosts. This led to the understanding that the relationship between prices and bookings is a central aspect to the host experience. Extremely high prices, relative to a listing’s specifications, should generally lead to very low booking rates, whilst too high booking rates could mean that the price is too low. In either scenario, the team assumes that the overall net profits are suboptimal. A simplifying assumption made for the subsequent analysis is that *a host with experience in a rental market will be able to identify the optimal balance, whilst newcomers may struggle for an extended period.*

Continuing on from the previous discussion on prices and booking rates, our initial hypothesis was that the AirBnB industry in NYC is divided into a ”short-term” and ”medium-term” rental market which can be distinguished by their respective price/booking equilibria. This distinction was inspired by the research we conducted as part of understanding the business case. Since the regulations were put in place partly to prevent short-term rentals, it made sense to explore how these listings differ from others with longer minimum night stay requirements.

Since the availability of a listing for the next 365 days is given directly in the Inside Airbnb datasets, this feature functions as an inverse proxy for the booking rate. We wanted to use this availability feature along with the price feature in order explore how these two factors impact the presence of a ”short-term” or ”medium-term” market. Different ranges of prices and availabilities were grouped and classified as ”short-term”, ”mixed-term” or ”medium-term” based on the average minimum night stay.

Figure 10 clearly shows that listings with higher prices and higher availability typically correspond to the short-term market. Conversely, the listings with lower availability and lower prices typically have higher minimum nights corresponding to a ”medium-term” market. The main insight derived from this analysis is the confirmation that the short-term and medium-term markets clearly differ in terms of their price/availability equilibria. The fact that short-term listings generally had higher prices made intuitive sense to the team, as these listings are typically booked in shorter advance and therefore expect a premium in return.

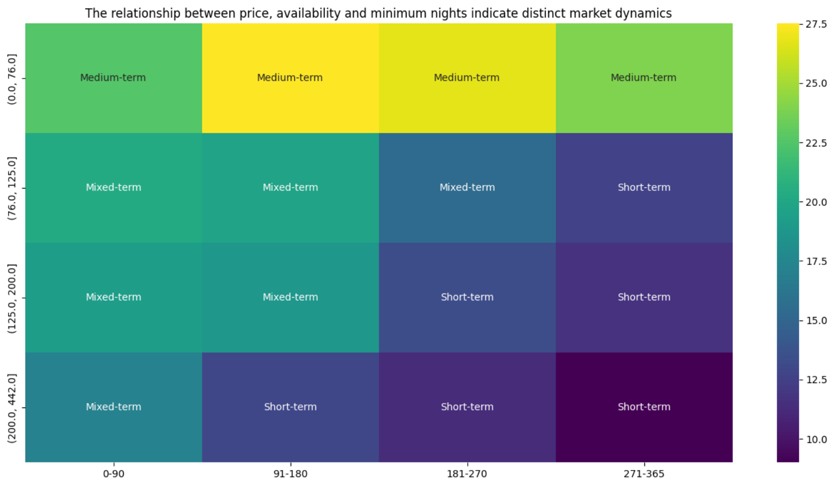


Figure 10: Various groupings of price and availability. The minimum night values are divided into three distinct categories. The y-axis shows price ranges and the x-axis shows availability ranges for the next 365 days. The colored bar on the right indicates the range of minimum nights explored.

4.2.3 Quantifying the Impact on Identified Host Segments

Having reviewed the price/availability equilibria for the two distinct rental markets in NYC, the next step is to examine exactly **how** the regulations have forced the displaced short-term hosts to adapt. Figure 11 shows the price/availability equilibria for two host groups; previous short-term hosts and veteran 30-day hosts. The main goal of this plot is to understand exactly how the price/availability relationship for the previous short-term hosts has changed since the new regulations were imposed. The prices and availabilities for the two host groups show distinct trends over the period of November to August, i.e. before the regulations. After this point, the short-term hosts are placed into the 30-day market due to the regulations. However, a key observation is that the price levels for the affected hosts remains relatively stable. Meanwhile, we see a marked increase in availability going from August to October.

Our initial insight from this figure is that many displaced hosts are choosing *to continue to operate from a short-term rental perspective in terms of pricing.* If the new 30-day hosts were sincerely attempting to adapt to the medium-term market, we would have expected to a see a much steeper decrease in the average price levels. If we now assume that the new 30-day hosts will receive negligible additional bookings as long as their prices remain too high compared to the rest of the 30-day market, then it is possible to quantify the lost revenues for these hosts for the next year. We find that the affected hosts in total will lose roughly **86 million dollars** in potential revenues for the next 365 days.

This projected revenue loss requires a more thorough explanation. In a perfect world, if all the affected hosts immediately adjusted their prices to the correct levels and received the same booking volumes as the established 30-day hosts, then they would on average gain 80 days of bookings (subtract the availability curve for new 30-day hosts from the curve for the veteran 30-day hosts in October) with an average price of 90 dollars per night (price curve for veteran 30-day hosts in October). This equates to 7200 dollars in revenues lost for a single host over the next 365 days, and when multiplying this loss by the 12’000 displaced hosts, we see a loss of roughly 86 million dollars. Naturally, this is a very rough estimate, but it helps to understand the massive danger the new regulations represent.

In the initial ”Ideation” phase for the project, we had assumed that any disruption from the regulations would dissipate quickly, and so we were more focused on how Airbnb could survive this transitionary phase as opposed to tackling it head-on. However, after quantifying the sheer scale of the potential losses and seeing that the new 30-day hosts are struggling to find the new optimal price, we now pivot to a more direct approach. The third objective will seek to determine the correct prices for the affected hosts through a price prediction model.

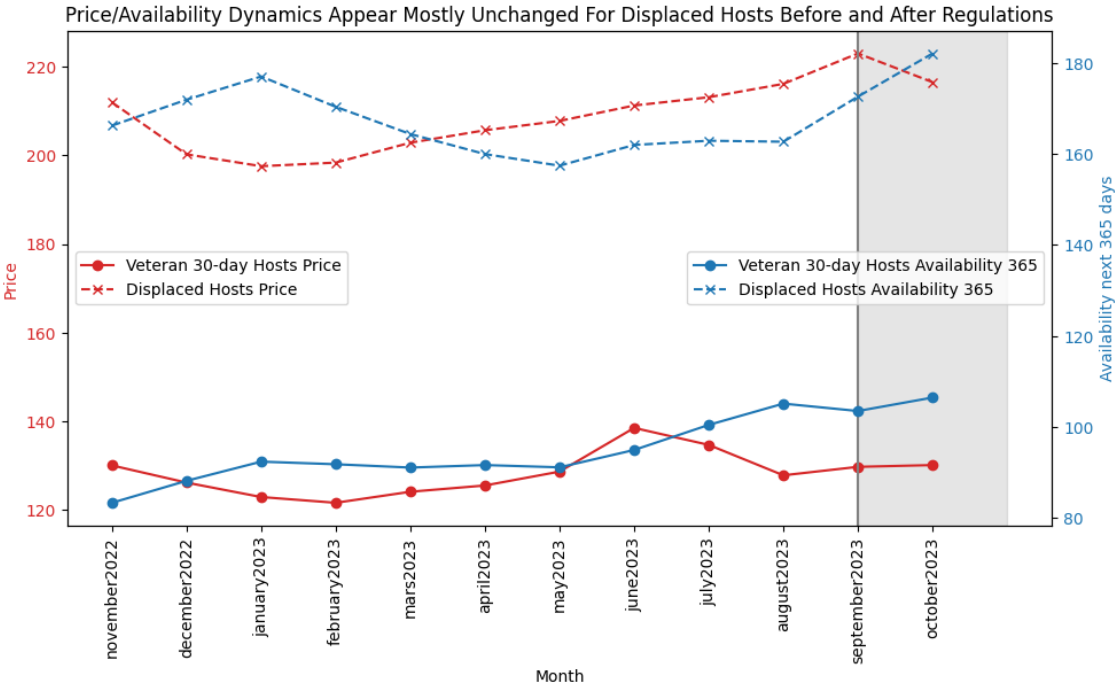


Figure 11: The plot is two-fold. The first part before the grey area indicates the short-term (dotted lines) and 30-day (smooth lines) price/availability dynamics before the new regulations. The grey area indicates the price/availability dynamics of veteran 30-day hosts and new 30-day hosts (AKA displaced short-term hosts) after the regulations.

4.3 Objective 3: Assist new medium-term hosts through specific pricing proposals

Following the confirmation that recent Airbnb regulations have significantly impacted the market by shifting many short-term rentals to medium-term offerings, we will proceed to develop a machine learning model. This model aims to aid these new medium-term hosts in setting appropriate prices for their listings.

The process will begin with the creation of training and prediction datasets, followed by the experimentation with various models. After selecting the most suitable model, we will refine it, interpret the results, and discuss their implications. The ultimate goal is to fulfill the objective of assisting new medium-term hosts through specific pricing proposals.

4.3.1 Preparing and Testing Models

4.3.1.1 Defining the prediction problem

Prior to model training and prediction, it is important to define which type of prediction problem we have. In data science, typical prediction problems are classified into categories such as classification, regression, and ranking. As our objective is to predict pricing, a continuous variable, it can be treated as a regression problem.

4.3.1.2 Working Set and Prediction Set

To effectively train and evaluate our machine learning model, we need to establish two distinct datasets: the Working Set and the Prediction Set.

Working Set
The Working Set is designated for both training and validating the model, typically divided into a training subset and a test subset. The training subset is used to fit the model, while the test subset is used to evaluate its performance. This division is critical for assessing the model's predictive power on unseen data. Our Working Set consists of listings that were offered for medium-term rental both before and after regulations started.

To ensure the model predicts accurately for the average rental, outliers were removed based on the accommodation capacity after segmenting the data by 'neighbourhood_group_cleansed' and 'room_type'. This step prevents the model from being influenced by atypical listings and improves its predictions for the general rental market.

Prediction Set
Conversely, the Prediction Set is intended for the model's application to generate actionable insights. It comprises the actual data on which the model's predictions will be applied, providing pricing guidance to users. This set includes listings that shifted to medium-term renting following the regulatory changes—those listed as short-term rentals in previous datasets, but are medium-term now.

4.3.1.3 Error Metric

Root Mean Squared Error (RMSE) was selected as the performance metric for its sensitivity to outliers. RMSE calculates the square root of the average squared differences between predicted values and actual values. This approach penalizes larger errors, which is desirable in price prediction as substantial deviations are more critical than smaller ones.

While Mean Absolute Error (MAE) was considered, it does not emphasize larger errors as heavily since it computes the average of absolute differences. Therefore, RMSE was deemed more appropriate for our objectives.

4.3.1.4 Testing Models

In machine learning, it is efficient to initially test a variety of models to identify the most promising one(s) for further tuning. We began by evaluating several commonly used models:

- Linear Regression
- Ridge
- Lasso
- Decision Tree Reg
- Random Forest Reg
- Gradient Boost Reg
- SVR
- Neural Network

We assessed the performance of these models using the Root Mean Square Error (RMSE) after training and testing them on a split of the working set. The RMSE results are summarized in the table below:

Model	**RMSE**
Linear Regression	0.516
Ridge	0.516
Lasso	0.569
Decision Tree Regressor	0.518
Random Forest Regressor	0.493
Gradient Boosting Regressor	0.495
SVR	0.929
Neural Network	0.579

Table 3: RMSE of Different Models

We also considered the ratio of predicted values to actual values, denoted as $y_{\text{pred}}/y_{\text{val}}$. This metric, while unconventional, provides a straightforward visualization of model performance. An accurate model would yield a ratio of 1, indicating perfect prediction. The corresponding box plot is shown below:

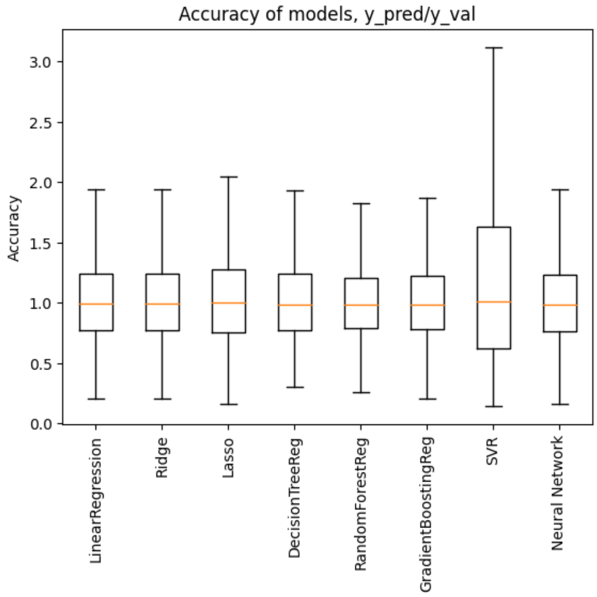


Figure 12: Box plot showing the ratio of predicted to actual values for various models.

Both the box plot, and most importantly, the RMSE values indicate that the Random Forest model outperforms the others, showing a median ratio close to 1 and a lower RMSE. Consequently, we selected the Random Forest Regressor, described in section 3.2, for further optimization.

4.3.1.5 Random Forest Regressor Optimization

K-Fold Cross-Val
We employed k-fold cross-validation to assess our model's performance, benefiting from its ability to validate and train on all data points, thereby reducing reducing bias. K-fold

cross-validation tends to enhance model accuracy but comes at the cost of significantly longer run and training times. However, as the goal was to predict as precisely as possible, we used it.

Hyperparameter optimization
Hyperparameter optimization was utilized to determine the best settings for our Random Forest model. This process involves selecting the most effective hyperparameters, which are not learned from data but are important for model training. We used Grid Search (GridCV) for this purpose, which exhaustively searches through a specified range of hyperparameter values to find the combination that results in the highest performance.

The optimal hyperparameters were as follows:

```
{ 'bootstrap': False, 'max_depth': 20, 'max_features': 'sqrt',
'min_samples_leaf': 2, 'min_samples_split': 5, 'n_estimators': 300 }
```

These parameters significantly improved our model's predictive accuracy.

4.3.2 Results and Interpretations

4.3.2.1 Results

Once the model was applied to the prediction set, price predictions were generated. These data-driven price predictions were then used to assess the pricing decisions of newly displaced medium-term hosts. The predictions were compared with the actual prices set by new medium-term hosts. The comparison used the relative difference, calculated as the ratio of the predicted to the actual price. The analysis indicated that the model's average prediction is 0.8 times the actual price set by the hosts.

$$\frac{y_{\text{predicted}}}{y_{\text{actual}}} = 0.8205$$

This ratio suggests that hosts transitioning to a 30-day rental model may be setting their prices higher than what the market data would support. To illustrate the distribution of our model's predictive performance, we have rendered the ratio of predicted to actual rent prices in a violin plot, as shown in Figure 13.

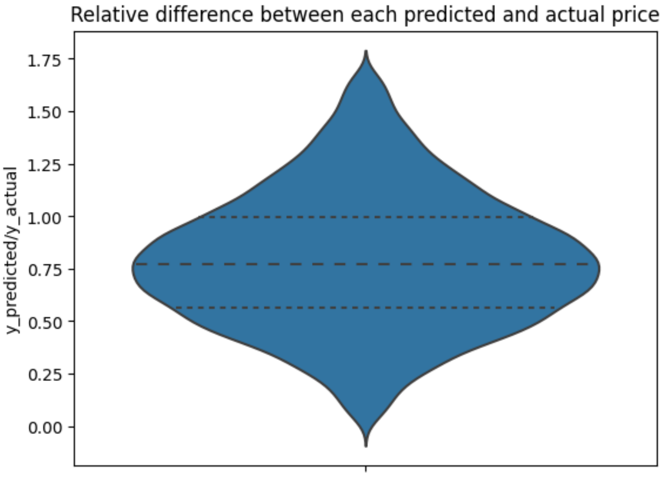


Figure 13: Violin plot of the relative difference between predicted and actual prices.

Violin plots are particularly useful for such an analysis because they show the density of the data at different values, providing a deeper understanding of the distribution of the predictions. From the violin plot, we can observe that most predictions cluster around the 0.8 ratio, with fewer occurrences of extreme over- or under-prediction. This visual representation reinforces the finding that there is a general tendency among hosts to overprice their rentals when compared to the model's predictions.

4.3.2.2 Feature Importance

An integral part of our project's objective is not just to predict new rental prices post-regulation but also to maintain model interpretability. The ability to explain our price predictions is essential for justifying the model's outcomes to customers. Therefore, understanding which features the model prioritizes when predicting prices is critical. We utilize the feature importance functionality inherent in the random forest algorithm to achieve this.

Feature importance in a random forest model provides a measure of the impact each feature has on the accuracy of the predictions. Essentially, it indicates how valuable each feature is in the construction of the decision trees within the model. The higher the feature importance, the more significant the feature is in predicting the target variable, in this case, the price.

Feature importance is depicted in Figure 14.

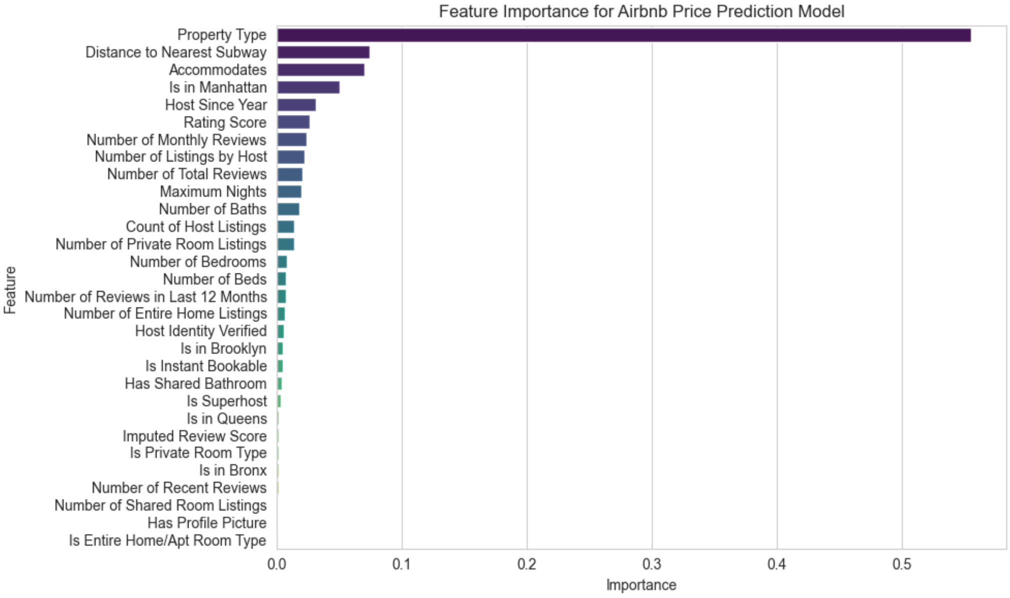


Figure 14: Feature importance from the random forest model. Note that Staten Island, while a borough, is not a feature as it is implicitly represented by the four other one-hot encoded boroughs.

It is important to note that a feature's high importance does not necessarily mean that its presence will increase the predicted price. Instead, it indicates that the feature significantly influences the model's price predictions, regardless of whether the effect is to increase or decrease the price. Features such as proximity to subway stations, accommodation capacity, and review quality logically influences pricing, aligning with our understanding of rental price determinants. Next, we will analyze some of these factors closer.

Property Type

The feature importance analysis identifies property type as the most crucial feature in determining price. Our dataset encompasses approximately fifty distinct property types, each falling into its own pricing category. To illustrate the impact of different property types on price, we utilized a SHAP plot. SHAP plots are valuable because they quantify the contribution of each feature to the prediction of the model, allowing us to understand how individual property types and their values sway the model's output.

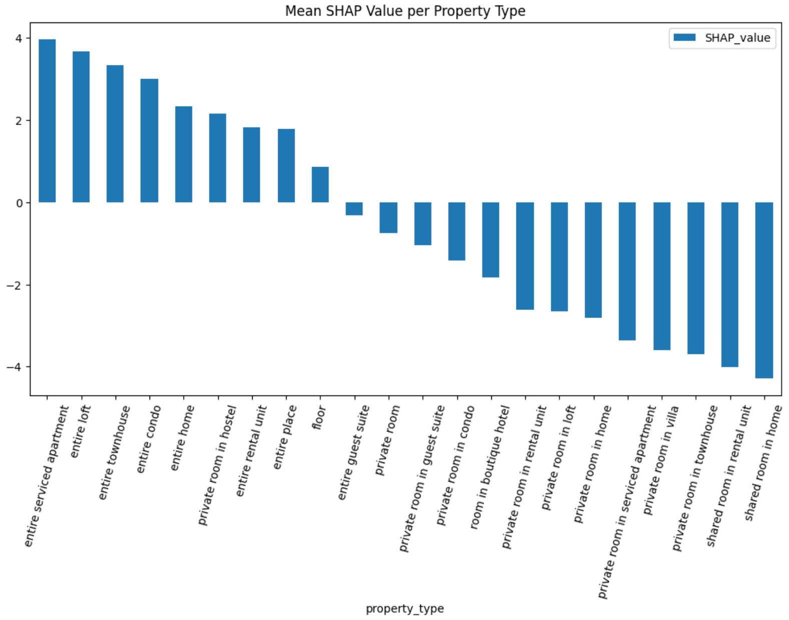


Figure 15: SHAP values depicting the influence of different property types on price predictions.

The SHAP plot demonstrates that property types have a significant influence on price. Listings described as 'entire home' or 'entire loft' or similar are associated with a positive effect on the price, while those labeled as 'private room' or 'shared room' tend to decrease it. This distinction aligns with intuitive expectations regarding the value assigned to privacy and space in rental properties.

Pricing Based on Borough

Of particular interest is the one-hot encoded variable for Manhattan. This feature stands out for its relative importance in the model, compared to the other one hot encoded features for Queens, Bronx and Brooklyn. This can be rationalized by examining the average prices across different neighborhoods. As depicted in the figures below, there is a substantial divergence in price for listings located in Manhattan compared to other boroughs.

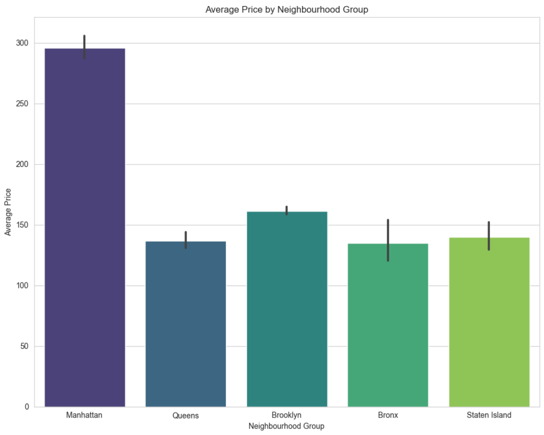


Figure 16: Comparison of average rental prices by neighborhood group, highlighting the disparity with Manhattan.

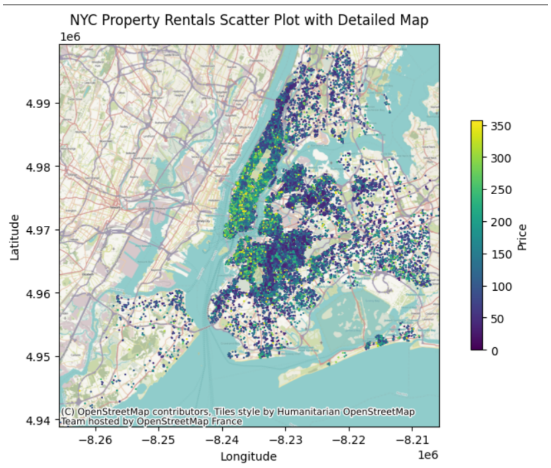

NYC Property Rentals Scatter Plot with Detailed Map
1e6
Latitude
4.99
4.98
4.97
4.96
4.95
4.94
−8.26
−8.25
−8.24
−8.23
−8.22
−8.21
Longitude
1e6
(C) OpenStreetMap contributors, Tiles style by Humanitarian OpenStreetMap Team hosted by OpenStreetMap France
Price
350
300
250
200
150
100
50
0

Figure 17: Scatterplot overlay on New York map showing rental prices distribution.

The plots illustrate that the average price in Manhattan is markedly higher than in other boroughs, which suggests that location within Manhattan is a major determinant of price. In contrast, the prices in other boroughs are relatively homogeneous. This explains why the model assigns such a high importance to the Manhattan variable—it provides a significant distinction in predicting price. In contrast, it does not gain much additional differentiation whenever the rental is located in other boroughs.

Superhost
The ‘Superhost’ status is heavily promoted by Airbnb as a mark of quality and reliability (Airbnb, 2023a), suggesting it should be a significant factor in pricing. However, our model showed that the ‘Superhost’ feature had a very low importance in determining price. This led us to hypothesize that the advantage of being a superhost might instead be more frequent bookings. To investigate this, we visualized the booking rates for ‘Superhosts’ against those without the status.

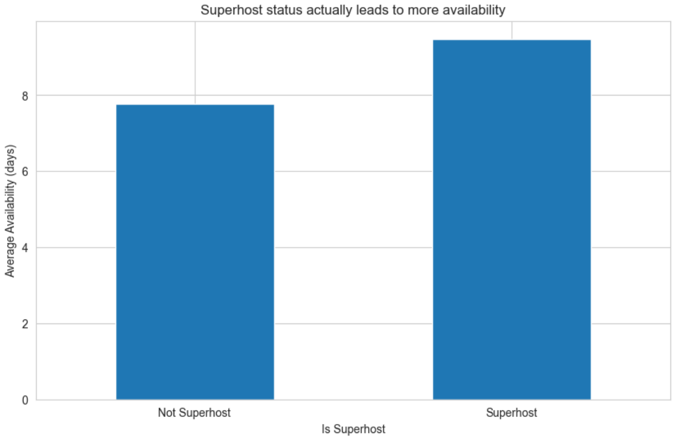


Figure 18: Comparison of availability for the next 30 days between Superhosts and non-Superhosts.

Contrary to expectations, the data showed that ‘Superhosts’ had a lower booking rate. This data challenges the assumption that ‘Superhost’ status directly correlates with increased demand.

4.4 The Iterative Process with CRISP-DM

Throughout the project, the iterative nature of the CRISP-DM framework made us continually reevaluate our business understanding, data understanding and the business value derived from the analysis performed for the selected objectives. In general, this iterative process continues on to the deployment and monitoring phases as well. However, since these final phases are never physically implemented, the iterative nature of the project mainly concerns the first five phases of the framework. This section will aim to describe some key experiences and insights from the iterative process.

4.4.1 Business Understanding

The business understanding was initially limited to a rudimentary understanding of how the Airbnb process works for guests, and the fact that Airbnb's business model is based on collecting a percentage fee from the bookings. However, the team had zero experience from a host perspective. One of the initial ideas brainstormed for the project was to develop a general pricing model as seen from Table 2. This candidate objective was assigned high business value initially. However, the first dive into the Inside Airbnb datasets showed that the data was heavily segmented by geographical location. A general pricing prediction model would therefore either need to be specialized for a particular area, reducing the business value, or generalized by merging hundreds of datasets, significantly reducing the feasibility. Thanks to this initial data understanding, we were able to quickly discard this objective.

After arbitrarily choosing to look at the NYC area, we discovered the overwhelming number of hosts with a minimum night stay of 30. This helped us to uncover the new regulations, which consequently led to the creation of several new candidate objectives. These instances show that the business understanding naturally developed alongside the data understanding, and was in part due to the strong emphasis on an iterative process through CRISP-DM. If we had maintained our original business understanding using a waterfall methodology Hotz, 2023b, the business value derived from the project would undoubtedly be far lower. The business understanding was naturally extended during later CRISP-DM stages as well, but this is covered in more detail under Section 4.4.5.

4.4.2 Data Understanding

The vast majority of our data understanding came from unexpected findings in later CRISP-DM stages. For instance, it was first during the modeling phase that the team realised how seasonality may impact the prices derived from the prediction model. This made us finally grasp how simply comparing prices and availability between August and October would not be sufficient for understanding the impact of the regulations. Hence, Figure 11 was modified to include all available months going back to November 2022.

The price column itself had been taken for granted in the initial data understanding. It was only during the data preparation that multiple members pointed out how it was unclear exactly how the price was derived. Revisiting this point enhanced our data understanding when we realized that the prices were most likely derived from the price per night of a monthly stay. This consequently made us consider the impact of weekly or monthly discounts on the price/availability dynamics of the short-term and medium-term rental markets, which helped us to understand why medium-term listings typically had lower prices as observed in Figure 10.

4.4.3 Data Preparation

For the first two objectives, the data preparation had been severely limited. The majority of the preprocessing steps explained in the Methods section were designed with the price prediction model in mind. The reason for this was that the first two objectives encompassed features which had zero missing values and did not require any corrections, such as the price and availability columns. As a result, the data preparation for the first two objectives did not require any noteworthy iterations.

The data preparation became far more intricate after the first evaluation phase, in which the price prediction objective was added. This objective required all categorical features to be encoded, and was initially done through one-hot encoding. However, during the subsequent modelling stage, it became apparent that the number of one-hot encoded features would make the models too computationally expensive to run. We therefore returned to the data preparation stage in order to remedy this problem, for instance by converting the one-hot encoded amenities feature into a count of amenities for each listing.

4.4.4 Modeling

The first modeling iteration was primarily concerned with establishing a baseline and comparing it to the more complex prediction models in order to evaluate their output. The original baseline was simply the mean price over all observations in the October dataset. However, we later revisited the price column to increase our data understanding, and subsequently discovered that the price distribution was heavily skewed as seen from 6. We therefore replaced this baseline with the median, and also log-transformed the prices, which gave stronger results for both the baseline and the prediction models.

We had originally disregarded model interpretability when ranking the candidate models. However, the analysis we conducted based on Figure 11 enhanced our initial business understanding. It became apparent to us that the displaced hosts were stubbornly refusing to adapt to the new price levels. We therefore began to doubt that a price prediction model without any context would be convincing enough to hosts. This led to another iteration, in which an additional requirement for the prediction model was that the price suggestion could be interpreted from the effects of the various listing features.

This final iteration made us reconsider the choice of the random forest regressor, as simpler models such as the decision tree or linear regression are more interpretable. However, given the random forest regressor's superior RMSE score and feature importance functionality, this model remained the preferred option. Hence, We concluded that feature importance provides sufficient context to the displaced hosts, and should hopefully convince more of them to adjust their prices to the new medium-term equilibrium.

4.4.5 Evaluation

In the evaluation phase, we wanted to review the entire process leading up to this stage in order to assert whether or not our insights from the analysis could be transformed to our business objectives. This evaluation process is divided into two main iterations, in which the evaluation phase during the first iteration resulted in a major upheaval for the project.

Iteration 1 was focused on discovering trends in the data which indicated both a presence and impact from the new regulations in NYC. This was first done by visualizing the changing trend in the average number of minimum nights before and after the inclusion of the new regulations, as seen in Figure 7. This confirmed the presence of the new regulations, thereby completing Objective 1. This objective may initially seem irrelevant in terms of contributing to the original business goal, given in 1.4, however it is always important to establish the presence of a business problem before diving in to solve it. If the analysis for this objective had uncovered that very few hosts had been moved to the 30-day market, then the best choice for Airbnb may simply have been to resume the status-quo. Alas, 7 indicates that a vast amount of hosts have been affected, which means that Airbnb will need to take a more active stance in order to handle the changes.

Further analysis focused on who would feel the impact of the regulations by filtering on number of listings and property type. This gave us the vital insight that many regular households, who may only seek to rent out a single room, are likely to experience the consequences of the regulations. Figure 10 goes on to show how the affected hosts are failing to adopt the new price/availability equilibrium for the medium-term market, and further examination led to the understanding that each of these hosts are potentially losing thousands of dollars given the current suboptimality. We deem these insights sufficient for Objective 2. The insights from this objective will help Airbnb to better target the affected hosts in order to retain them on the platform, as they now know exactly who the affected hosts are and understand their characteristics and the potential revenues they are losing. This will allow Airbnb to more effectively handle the regulatory changes by enabling them to clearly communicate with the affected hosts regarding their concerns.

The team never reached the point of evaluating Objective 4 during the first iteration, as the evaluation of the first two objectives led to a drastic change in our understanding of the business problem. We had originally assigned Objective 4 far higher business value than Objective 3, as seen from Figure 1. This was due to our initial business understanding, which had made us believe that the effects from the regulations were temporary. Our insights from Figure 11 initially strengthened this belief during iteration 1. We interpreted the lack of change in the prices of displaced hosts for October to mean that many hosts were confident in regaining their short-term status quickly, and were therefore choosing to ignore the dynamics of the medium-term market.

However, it soon became clear that we had misinterpreted the findings from Figure 11. This revelation had its origin in the evaluation of Objective 1, when a team member noted that the effect of the regulations on the number of 30-day hosts appeared to grow from September to October based on Figure 7. This change was not anticipated, as we had expected to observe a decreasing trend in 30-day hosts for October as short-term hosts began to regain their status. These findings led to an ad-hoc review of the new regulations, which subsequently resulted in a significant change in our understanding of the business problem. We discovered that the new regulations impose clear restrictions on renting out short-term without the host being present(Shirshikov, 2023). This gave us an entirely new perspective on the business problem, and we began to view the regulations as a paradigm shift for the displaced hosts as opposed to a temporary annoyance. The evaluation of Objectives 1 and 2 hence altered the trajectory of the subsequent analysis, which eventually led into the second iteration of this phase focused on Objective 3.

Iteration 2 of the evaluation phase was mainly concerned with the final objective of the data analysis, which had been changed from Objective 4 to Objective 3 after the findings from the previous evaluation phase. The new objective was to construct a pricing model which would help the the displaced hosts quickly adapt to the new market and converge on the same booking rates and prices as the existing 30-day hosts. This would in turn lead to increased revenues for Airbnb, as their service fees directly relate to how well the price/availability ratio is maintained for hosts on their platform. This marked the final significant change in our business understanding of the project.

As discussed previously, the majority of the data preparation was performed in relation to the price prediction modeling. In order to ensure the reliability of the pipeline, all relevant features for the price predictions were cleaned of missing values as explained in Section 3.3. Outliers were handled both through the removal of atypical listings and the RMSE metric, as explained in Section 4.3.1.2.

It was imperative that the prediction model related to the objective was properly evaluated with cross validation and the previously justified RMSE metric. Furthermore, we also needed to ensure that the model output made intuitive sense given our evolved business understanding. The best performing model for all iterations was the random forest regressor. During the second evaluation phase, we initially found that this model on average proposed a 10% decrease in price for the displaced hosts. However, Figure 11 shows that the prices of the displaced hosts are nearly double that of the veteran 30-day hosts for October. Hence, our business understanding and data understanding did not align with such a modest price decrease. We therefore performed a new round of modeling, which resulted in the inclusion of the subway feature and the removal of multiple highly correlated ratings features, as discussed in 3.3.

This resulted in the complete model, which has a further improved RMSE score and recommends on average a 20% decrease in price. Though these prices are still far above the levels of the veteran 30-day hosts, we explored the feature importance of this model, depicted in 4.3.2.2, and found that the model's pricing explanations were logical. Therefore, we accept the output of the model, and further deemed the price corrections sufficient for helping hosts to quickly reach the 30-day market equilibrium. Completing this final objective contributes towards the overall business goal. Specifically, the insights about correct price levels for the affected hosts will help them to achieve a new price/availability equilibrium, which in turn will prevent the loss of revenues for both these hosts and Airbnb.

4.5 Shortcomings of Methods Used

While our model provides valuable insights, there are inherent limitations in our approach that must be acknowledged.

Model Limitations The Random Forest Regressor, despite its robustness, can struggle with extrapolation outside the range of the training data. In a dynamic market like NYC's rental space, where prices can be influenced by sudden and unpredictable factors, the model may not capture future trends not present in the past data. Additionally, Random Forests can be computationally expensive, which may not align with the need for rapid price predictions for new hosts.

As discussed previously, the random forest regressor is somewhat less intepretable than some of the alternatives. A linear regression model could have given hosts an exact explanation of how the pricing was done by using the regression coefficients. Feature importance from the random forest regressor does indicate what has been most influential in the price prediction, but does not indicate exact magnitude or whether a feature has a positive or negative price impact. However, these downsides were partially compensated for by the use of Shap plots and other visualization tools in order to better understand the relationships to the price.

Feature Selection and Engineering Our feature importance analysis heavily relies on historical data. The relationships and trends inferred may not hold in a post-regulation landscape, especially since new regulations can alter the factors tenants value. For example, the importance of proximity to public transportation may change if remote work becomes more prevalent. Moreover, features like reviews and property type are subject to biases and self-reported data which can introduce noise into the model.

There are several features not present in the available data which could have been valuable for the prediction model. An estimate of the square footage of all listings would help to distinguish between listings with the same property or room type, thereby increasing the accuracy of the price predictions. Another useful feature would be more detailed pricing features. At the present, there is only one price feature which is seemingly derived from the average price for a night over a given period of time. However, there is currently no way to isolate weekly or monthly discounts from this feature, thus limiting the ability to accurately distinguish between the pricing strategies of short-term and medium-term listings. We would generally expect to more frequently observe the use of discount strategies for longer-term hosting, but this has yet to be confirmed. If a 'pricing strategy' feature was available, it could be incorporated as a part of the pricing suggestion for the displaced short-term hosts who may not be acquainted with this functionality. Given all of the above data-related issues, it would be prudent in the future to source the monthly data directly from AirBnB instead of using an open-source platform which is not directly affiliated with the firm.

Dataset Constraints The dataset represents a snapshot of the market before and after the regulation change, but it does not capture medium-term effects. It also excludes unlisted properties and those that may have been delisted due to the regulations, potentially skewing the price predictions. Another concern is the treatment of outliers, where the removal of extreme values may oversimplify the complexity of the market.

Assumptions in Model A fundamental assumption in our modeling approach is that the prices in the training set are correctly set for medium-term rentals. This assumption does not account for the possibility that even before the regulation changes, some listings could have been mispriced or influenced by factors not captured in the dataset.

Additionally, we assumed that the rental market could absorb the rapid increase in the supply of medium-term rentals without significantly affecting prices. Specifically, we did not model the potential for a market flood leading to a decrease in prices due to the sudden doubling of medium-term rental listings. This assumption overlooks the elasticity of demand and the complex dynamics of the housing market, which could lead to a significant reduction in price levels when supply increases sharply and unexpectedly.

Part 5

Deployment and Recommendations

5.1 Recommendations

Based on the previously stated objectives and analysis performed, we have created recommendations for Airbnb. These include the integration of the pricing algorithm, educational and awareness increasing efforts towards hosts, and actions for engaging in lobbying activities against new regulations.

5.1.1 Integrate our pricing optimization algorithm into the Airbnb application, with additional functionality for explaining the suggested price

According to the results in Figure 11, new 30-days hosts price their properties too high. To help them more accurately price their listings and reduce their adaptation phase in the new market, we advise Airbnb to integrate the developed pricing algorithm into the Airbnb application. It should be clearly displayed and easily accessible in the user interface for publishing and editing of listings. Furthermore, we also recommend Airbnb to supplement the price suggestions with detailed explanations of the prices, as it is likely to increase users' trust in the price suggestions. An example of an explanation can be: "similar properties for two people in this area with a shared bathroom are usually priced 100 dollars less".

5.1.2 Send out newsletters to new 30-day hosts explaining the medium-term rental market, with advertisement of the new pricing algorithm

The recent regulatory changes represent a black swan event, leading to many new medium-term landlords overpricing their listings. It is crucial for Airbnb to proactively address this issue. We recommend dispatching a newsletter that educates hosts on the medium-term rental market changes, the influence on various property types, and the introduction of a new pricing algorithm to assist in accurate pricing.

5.1.3 Further investigate the effects of the regulations and use insights to lobby for regulatory change

50% of hosts impacted by the new regulations possess only a single listing on Airbnb, as depicted in Figure 8. This suggests that the majority are individual hosts rather than large rental enterprises, the intended target of the regulations. It is recommended that Airbnb conduct a detailed analysis of the affected individuals to map out the regulations' scope. Such insights could be crucial for advocating modifications to the regulations to ensure they more accurately target their intended subjects, especially in other cities considering similar measures.

5.1.4 Create and distribute automated personalized reports on property performance

It is important for Airbnb that the properties on their platform are correctly priced over time. Therefore we suggest implementing automated personalized reports on property performance, akin to Spotify's user insights provided by 'Spotify Wrapped' (Naomi, 2022). The reports should be sent out to hosts semi-annually and include relevant statistics and comparisons to other similar properties, in addition to pricing suggestions and estimations of potential revenue increases. This effort is likely to serve as a motivator for hosts to adjust their prices in accordance to the proposed prices, which benefits Airbnb as well. To make the process easy and streamlined, the total process of creation and distribution should be automated.

5.1.5 Further enhancements to the current pricing algorithm

After the deployment of the model, further improvements should be the focus area. Examples of such enhancements can include continuously refreshing the model with new data to accurately reflect the market, predicting pricing trends, incorporating seasonal variations, and in-depth segmentation analyses to detect price sensitivities among different guest demographics.

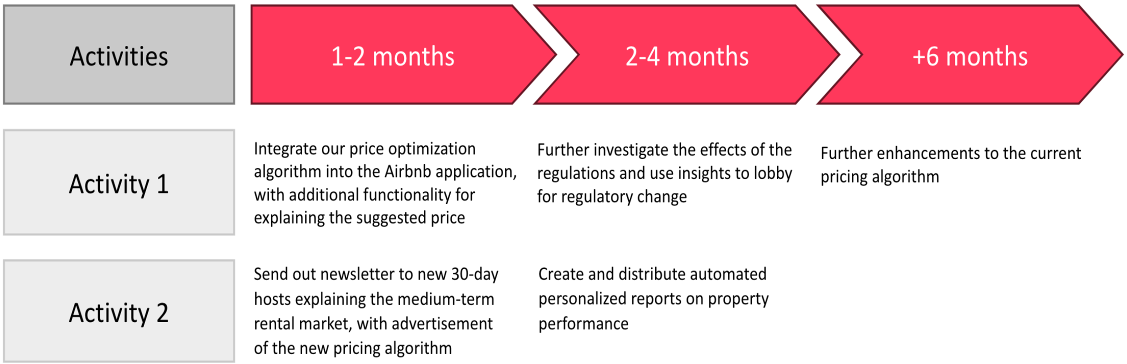
5.1.6 Effect on stakeholders

For the recommendations, number 1, 2, 4, and 5, are suggested to help hosts price their properties correctly. The primary stakeholders for these recommendations are new 30-days hosts, Airbnb shareholders, and renters on Airbnb. Hosts will benefit from an increase in revenues due to better balance of price and booking rates. Consequently, shareholders will benefit as the overall revenues of the platform increases. Renters on the platform will benefit due to the increased supply of correctly priced properties.

For recommendation 3, the expected stakeholders affected are regulatory authorities, the tourism industry, hosts across the world, and Airbnb shareholders. Global regulatory authorities will be affected by the increased resistance and lobbying activities against further regulations. On the other hand, the tourism industry, global hosts, and Airbnb shareholders are likely to benefit from the lobbying activities as these stakeholders stand to benefit from less regulations.

5.2 Implementation plan

For Airbnb to effectively deploy the recommendations in a timely manner, we have proposed an implementation plan, as shown in Figure 19. The plan is created based on the recommendations' urgencies and interdependencies. The very first recommendation is to integrate the pricing algorithm into the Airbnb platform, as it addresses the core problem of pricing. Due to the regulations already being in effect, we consider this recommendation urgent, and suggest Airbnb to assign significant resources for it. Additionally, for the first phase, we recommend an educational newsletter promoting the pricing algorithm. The reasoning for this is to increase hosts' knowledge and trust in the system. In the second phase we recommend to perform and publish an analysis on how regular people are affected, which can be used for lobbying activities. As stricter regulations is a concern in several cities, this recommendation is considered relatively time-critical. Phase two also includes a recommendation for semi-annual personal newsletters on properties' performance. This recommendation is natural to put in phase two as it is an effort to ensure correct pricing levels over time. Finally, in the last stage when the model up and running, the focus should move to further improvements of the model.



Activities	1-2 months	2-4 months	+6 months
Activity 1	Integrate our price optimization algorithm into the Airbnb application, with additional functionality for explaining the suggested price	Further investigate the effects of the regulations and use insights to lobby for regulatory change	Further enhancements to the current pricing algorithm
Activity 2	Send out newsletter to new 30-day hosts explaining the medium-term rental market, with advertisement of the new pricing algorithm	Create and distribute automated personalized reports on property performance	

Figure 19: Implementation plan for recommendations

5.3 Future analysis

Currently, the training data used is rental data for experienced 30-days hosts before the enforcement of the regulations. Therefore, a major assumption is that these hosts priced themselves correctly. Additionally, the rental market is currently facing several uncertainties, including increased interest rates and inflation. Also, there is a significant possibility of the medium-term rental market becoming even more destabilized with the increase in supply of medium-term properties as a result of the regulations. Due to these limitations, we recommend Airbnb to continuously integrate recent data into the model to make it more responsive to changing dynamics in the market.

Furthermore, the model can be enhanced by incorporating elements such as seasonality and in-depth market segmentation. The seasonality aspect includes analyzing price variations throughout the year, for example changes during holidays or summer. Regarding market segmentation, the model already considers location and type of property, but a more in-depth analysis is likely to be beneficial . This analysis can include neighboorhoods, comparing apartments, houses, and unique accommodations, and preferences and price sensitivities within groups such as business guests, solo travellers, and families.

There are several limitations associated with the choice of data source, namely Inside Airbnb. The monthly datasets for NYC only go back to November 2022, which makes it impossible to view trends spanning multiple years. For instance, Figure 11 could have been extended for multiple years in order to show how the discrepancy between short-term hosts and medium-term hosts is an established trend. There are also a multitude of features in the dataset which could be valuable, had it not been for the significance presence of randomly missing values as discussed in the Methods section. One such example is the ”License” column, which in all likelihood is meant to highlight whether or not the host has the appropriate licensing for a listing. Unfortunately, the October dataset only has missing values for this feature, which is why it was immediately dropped during the initial data exploration. If it was possible to separate out the licensed hosts from the unlicensed hosts, then the price predictions could be further concentrated in on the unlicensed hosts which are likely to remain in the medium-term rental market. For future analysis, an attempt should be made to receive the listings data directly from Airbnb, as opposed to a third-party intermediary.

Part 6

Monitoring and Maintenance

6.1 Key Performance Indicators

Effective monitoring of the new algorithm is important for achieving its intended outcomes. For this, key performance indicators (KPIs) are used. These KPIs are divided into two categories; model performance and business performance, as shown in Table 4.

Collectively, these KPIs provide a holistic evaluation of the solution. Adoptation rate and reading rate give an indication of hosts' trust in the model, and therefore functions as a proxy for the accuracy. The technical performance KPI measures whether the model meets technical benchmarks. Furthermore, the robustness KPI is an important indicator of the model's ability to handle a diverse rental market like . For the business KPIs,

Model performance KPIs	
Adoption rate	Percentage of hosts using the pricing tool yearly
Alarm signal	Adaptation rate below 50% among new 30-days hosts
Action plan	Increase outreach and education through intensified marketing efforts, regular user feedback collection, and detailed explanations about the tool's functionality
Responsible team	Marketing team and customer support
Reading rate	Percentage of hosts reading the semi-annual newsletter
Alarm signal	Reading rate below 30%
Action plan	Overhaul the newsletter by including more tailored content and testing new types of content
Responsible team	Marketing team
Technical performance	Measures of latency and uptime for the model
Alarm signal	More than 10% of recommendations exceeding 5 seconds to process or an uptime below 99.5%
Action plan	Enhance system reliability with regular maintenance, infrastructure upgrades, and expanded cloud resources
Responsible team	IT support team
Model robustness	Ability to predict reliably across diverse properties
Alarm signal	Significant fluctuations in performance metrics for different areas, property types etc.
Action plan	Improve model's effectiveness by performing an analysis to pinpoint the root causes of performance irregularities and expand the dataset with more instances where the model currently has difficulties
Responsible team	Data science team
Business performance KPIs	
Booking rate	Percentage of days new medium-term properties are booked
Alarm signal	A 5% lower booking rate for new 30-days hosts using the tool compared to experienced 30-days hosts
Action plan	Conduct an in-depth examination of the pricing algorithm to pinpoint areas of pricing misalignment and make necessary adjustments to the model
Responsible team	Data science team
Revenue generation	Average revenue per day of new medium-term properties
Alarm signal	Average daily revenue 10% lower for new 30-days hosts compared to experienced medium-term hosts
Action plan	Perform detailed review of pricing algorithm focusing on uncovering differences among new and experienced medium-term hosts, such as suitability for medium-term rental, and adjust model accordingly
Responsible team	Data science team
Host retention rate	Churn rate of new 30-days hosts
Alarm signal	5% rise in the churn rate of hosts, excluding those who depart the platform due to their inability to offer medium-term rentals
Action plan	Launch host retention initiative though conducting surveys for understanding causes of churn and increasing host support
Responsible team	Customer support team
Market share	Changes in market share
Alarm signal	5% decline in market share in house-sharing market
Action plan	Perform comprehensive competitive analysis of the house-sharing market and incorporate new data and trends reflecting current market dynamics into the model
Responsible team	Business development team and data science team

Table 4: KPIs for model performance and business performance

booking rate and revenue per day are important indicators of Airbnb's overall health and profitability. Additionally, booking rate also functions as a proxy for model accuracy, as deviations in booking rates between new and experienced medium-term hosts indicate wrong pricing level. Furthermore, host retention rate and market share are important for evaluating changes in platform satisfaction and competitive performance.

Together, the KPIs give an overview of the project's effectiveness, and will inform Airbnb if something is significantly deviating from the expected performance. Additionally, as an extra precaution, we advise Airbnb to perform a full review of the solution six months after implementation to evaluate its success.

6.2 Lessons learnt and feedback

This data science project has been a valuable learning experience for the group, and we are left with several insights that can benefit similar projects in the future. Our first lesson learnt is to determine early whether an interpretable model is necessary or not. By deciding on this, one can possibly quickly disregard black box models and therefore save significant time. Secondly, we learnt that it is important to conduct a thorough analysis, rather than quickly confirming that the trend in the dataset aligns with the initial hypothesis. In our case, we noticed the trend of increasing minimum nights early, but we were only able to understand the severity when we dived deeper into the data. A more in-depth analysis benefited us significantly because it led us to reevaluate our initial perspective, and we where then able to understand the real problem and address it at its core. When we created the model, we learnt that most users look at the large overall picture, rather than small details, when renting properties on Airbnb. Therefore, we recommend similar work in the future to focus on overarching features, instead of "quality of life" features such as amenities or thoroughly investigating specific reviews. Furthermore, if aiming to distinguish between short-term and medium-term rental markets, it is important to early identify the distinct dynamics in pricing and booking rates between the two markets. Lastly, a very important lesson learnt is the importance of data science project management. The combination of CRISP-DM and design thinking enabled the group to work in a structured and efficient manner throughout stages, and allowed us to easily go back and make adjustments without any time-consuming errors.

Bibliography

Airbnb. (2023a). *What is a superhost?* Retrieved 23rd November 2023, from https://www.airbnb.no/help/article/828

Airbnb. (2023b, November). *How airbnb works - Airbnb Help Center*. Retrieved 25th November 2023, from https://www.airbnb.com/help/topic/1378

Atanda, A. (2023). *Battle of Visuals: Matplotlib vs Seaborn in Data Science*. Retrieved 22nd November 2023, from https://www.linkedin.com/pulse/battle-visuals-matplotlib-vs-seaborn-data-science-adeoluwa-atanda

Chan, W. (2023). 'We're in a housing desert': A month in, is New York's Airbnb crackdown working? *The Guardian*. Retrieved 25th November 2023, from https://www.theguardian.com/us-news/2023/oct/23/new-york-airbnb-crackdown-rules-housing

Dam, R. F., & Siang, T. Y. (2022). *What is design thinking and why is it so popular?* Retrieved 12th November 2023, from https://www.interaction-design.org/literature/article/what-is-design-thinking-and-why-is-it-so-popular#what_are_the_5_phases_of_design_thinking?-1

Hotz, N. (2023a). *What is crisp dm?* Retrieved 12th November 2023, from https://www.datascience-pm.com/crisp-dm-2/

Hotz, N. (2023b). *What is waterfall?* Retrieved 22nd November 2023, from https://www.datascience-pm.com/waterfall/

Inside Airbnb. (2023a). *About inside airbnb*. Retrieved 13th November 2023, from http://insideairbnb.com/about/

Inside Airbnb. (2023b). *Get the data*. Retrieved 16th October 2023, from http://insideairbnb.com/get-the-data/

Jupyter. (2023). *Jupyter*. Retrieved 22nd November 2023, from https://jupyter.org/

Naomi. (2022, November). Everything You Need To Know About 2022 Wrapped. Retrieved 25th November 2023, from https://newsroom.spotify.com/2022-11-30/everything-you-need-to-know-about-2022-wrapped/

New York Open Data. (2023). *Mta subway stations*. Retrieved 20th October 2023, from https://data.ny.gov/d/39hk-dx4f

NVIDIA. (2023). *Pandas*. Retrieved 22nd November 2023, from https://www.nvidia.com/en-us/glossary/data-science/pandas-python/

O'Sullivan, F. (2023). Airbnb Hosts Try to Evade City Regulations, From Copenhagen to Catalonia. *Bloomberg.com*. Retrieved 25th November 2023, from https://www.bloomberg.com/news/features/2023-08-02/cities-keep-trying-and-failing-to-regulate-airbnb-nasdaq-abnb

Pafitis, S. (2022, May). *What Makes Python a Great Pick for Data Analysis*. Retrieved 22nd November 2023, from https://www.linkedin.com/pulse/what-makes-python-great-pick-data-analysis-sotiris-pafitis-1f

Saltz, J. (2022). *Crisp-dm is still the most popular framework for executing data science projects*. Retrieved 12th November 2023, from https://www.datascience-pm.com/crisp-dm-still-most-popular/

Seya, H., & Shiroi, D. (2022). A comparison of residential apartment rent price predictions using a large data set: Kriging versus deep neural network. *Geographical Analysis*, *54*(2), 239–260. https://doi.org/https://doi.org/10.1111/gean.12283

Shah, K., Shah, H., Zantye, A., & Rao, M. (2021). Prediction of rental prices for apartments in brazil using regression techniques. *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, 01–07. https://doi.org/10.1109/ICCCNT51525.2021.9579796

Shirshikov. (2023). *A guide to airbnb and vacation rental regulations in new york state*. Retrieved 25th November 2023, from https://awning.com/post/new-york-state-short-term-rental-laws

Zaveri, M. (2023, September). *New york city's crackdown on airbnb is starting. here's what to expect*. Retrieved 14th November 2023, from https://www.nytimes.com/2023/09/05/nyregion/airbnb-regulations-nyc-housing.html

Appendix

A Host related features

host_id	A unique identifier of the host
host_since	The date when the host started to use Airbnb
host_location	The city or country the host is living in
host_is_superhost	A boolean saying whether the host is a superhost
host_total_listings_count	The total number of active and inactive listings for the host
host_neighbourhood	The neighbourhood a host is living in
host_has_profile_pic	A boolean saying whether a host has a profile picture
host_identity_verified	A boolean saying whether a host is verified
calculated_host_listings_count	The amount of listings the host has active currently
calculated_host_listings_count_entire_homes	The amount of listings for entire houses the host has currently
calculated_host_listings_count_private_rooms	The amount of listings for entire private rooms the host has currently
calculated_host_listings_count_shared_rooms	The amount of listings for shared rooms the host has currently
host_response_rate_	The rate at which a host responds to potential guests
host_acceptance_rate_	The rate at which a host accepts new bookings

Table 5: The host related features in the original dataset, which are not removed in the very first iteration.

B Listing related features

id	The unique identifier of the host
name	The name of the Airbnb listing
description	A description of the listing in plain text written by the host
neighbourhood_cleansed	The local neighbourhood of a listing, covers a relatively small area
neighbourhood_group_ cleansed	Main area where the listing is located
longitude, latitude	The coordinates of the listing
property_type	The type of the property, ranges from private room to tower
room_type	The type of the room, ranges from shared rooms to entire homes and hotel rooms
accommodates	How many people who can stay at the listing together
bathrooms_text	The number and type of bath(s)
bedrooms	The number of bedrooms
beds	The number of beds
amenities	A list consisting of amenities the listing provides, ranging from a hair-dryer to a pool
price	The price of a listing
minimum_nights	The minimum nights the listing has to be booked for a stay
maximum_nights	The maximum nights the listing can be booked for a stay
has_availability	A boolean saying whether the listing has any availability in the future
availability_30, **availability_60,** **availability_90,** **availability_365**	The number of days the listing is available for the next 30, 60, 90 and 365 days
number_of_reviews, **number_of_reviews_ltm,** **number_of_reviews_l30d**	The total number of reviews for the lifetime of the listing; last twelve months; last 130 days
instant_bookable	A listing which can be booked without pre-approval by the host
review_scores_rating	The overall average rating of a listing given by past tenants (1-5 stars)
review_scores_accuracy	The average impression by a tenant of how accurately described the listing is (1-5 stars)
review_scores_cleanliness	The average impression of how clean a listing is at arrival (1-5 stars)
review_scores_checkin	The average impression of the check-in process for the listing (1-5 stars)
review_scores_ communication	The average impression of how well the host communicates before, during and after the stay (1-5 stars)
review_scores_location	The average impression of the location of the listing (1-5 stars)
review_scores_value	The average impression of the listing's price-to-value (1-5 stars)

Table 6: The listing related features in the original dataset, which are not removed in the very first iteration.