ANUSHA KHAN

40390503

Title: Price Influencing Features of Airbnb Business in Boston: A Predictive Model for Rental Price Prediction of New Listings

WORDCOUNT: 8100

ABSTRACT

In the dynamic landscape of the modern sharing economy, Airbnb is recognized as a pioneering force that has redefined the lodging industry with its peer-to-peer model. The focus of the present study is placed on the role of machine learning algorithms in multiple facets of Airbnb's operations, including the optimization of pricing strategies, enhancement of consumer experiences, and provision of market insights. It is found that machine learning algorithms are employed to determine optimal listing prices by taking into account variables such as property type, geographical location, amenities, and historical reservation data. The precision of these algorithms is shown to assist consumers in making well-informed decisions regarding their choice of accommodation and budget planning. Insights into market shifts, demand cycles, and seasonal variations in pricing are generated by machine learning algorithms, which are shown to be instrumental for hosts, property managers, and investors in making data-driven decisions concerning property acquisitions and pricing approaches. For entities involved in property management, machine learning algorithms are demonstrated to facilitate the optimization of pricing strategies across multiple listings, thereby maximizing revenue potential through dynamic adjustments based on current market conditions and occupancy rates. The study also elucidates how machine learning algorithms contribute to strategic competitive positioning by enabling hosts to gauge their pricing competitiveness in relation to comparable listings, thereby allowing for informed rate adjustments aimed at attracting a greater number of bookings. Furthermore, machine learning algorithms are observed to play a significant role in risk assessment in relation to pricing decisions, offering hosts the capability to assess the potential benefits of reduced pricing during low-demand seasons to elevate occupancy rates and income. The segmentation of customer demographics through the analysis of historical reservation data and guest profiles by machine learning algorithms is explored, indicating the customization of pricing strategies to better suit various customer demographics. For property investors, machine learning algorithms are utilized for the estimation of potential revenue generation from Airbnb listings, thus aiding in property valuation and investment decisions. In localities where regulatory measures exist for short-term rentals, machine learning algorithms assist hosts in understanding and adhering to pertinent pricing rules and tax obligations. The study

concludes by providing a comprehensive examination of the myriad factors influencing Airbnb pricing, including but not limited to geographical location, property type and size, amenities, reviews, host experience, seasonality, and overarching market dynamics. This nuanced understanding is found to empower hosts in effectively managing their listings and maximizing revenue.

TABLE OF CONTENTS

INTRODUCTION

In the evolving landscape of the contemporary economy, the concept of ownership has undergone a profound transformation, particularly within the context of the Sharing Economy or Collaborative Consumption. The sharing economy aims to reinvent a business model redistributing goods and services creating a win-win situation, cultivating earning and spending opportunities on a local basis. The ownership of land represents a tangible asset with an intrinsic value that enables the generation of income via various means such as rentals, leases, and capital appreciation The sharing economy is centered on the goal of redistributing current possessions amongst the populace with the purpose of maximizing their utility(Guttentag, 2015). The sharing economy offers platforms that permit users to share their possessions with others, leading to the creation of fresh consumption patterns, with sharing not necessarily denoting a free-of-charge service. By framing land ownership within the framework of the Sharing Economy, this paper examines the possibilities of actively engaging in peer-to-peer transactions by property owners through sharing access to their land resources, facilitated by community-based online platforms. By doing so, property owners participate in collaborative consumption, a phenomenon where individuals can access and utilize land without full ownership. This demonstrates the adaptive nature of the Sharing Economy as it expands its reach beyond traditional goods and services, incorporating the shared use of tangible assets like land.

The concept of Airbnb is not formed by obliteration of old business models but actually reinvention or updating by use of technology and social media emergence.

The elasticity of supply and demand, which is a pivotal economic principle, has a critical influence on Airbnb's pricing strategy. In order to optimize revenue, the platform adapts prices in accordance with variations in supply, which refers to the number of available listings, and demand, which pertains to traveler preferences, events, and seasons. By utilizing machine learning algorithms, Airbnb can anticipate and respond to shifts in supply and demand. Consequently, during periods of high demand, prices rise, thereby stimulating more hosts to enlist their properties, leading to an expansion of supply and the maintenance of a balanced marketplace.

A research study highlights the importance of trust and reputation systems for online platforms, and this holds true for Airbnb as well. Given the nature of the transactions between strangers, trust-building mechanisms such as feedback and ratings are vital to establish credibility. In particular, Airbnb's review and rating system allows hosts and guests to assess each other, providing a mechanism to mitigate concerns related to sharing living spaces with unfamiliar individuals. The effectiveness of such a system enhances user confidence, which, in turn, fortifies platform adoption (Ma et al., 2017).

Airbnb utilizes a multitude of monetization strategies, notably encompassing commission-based fees, whereby hosts are required to pay a percentage of each booking to Airbnb, while guests may incur service fees. An additional means of optimizing revenue includes implementing surge pricing during peak seasons, a pricing model otherwise recognized as dynamic pricing. This particular approach effectively capitalizes on demand data, thus enabling Airbnb to adjust rates and augment revenue for both hosts and the company itself(Wirtz and Tang, 2016).

According to the Network Effects Theory proposed by Metcalfe in 1995, a network's value is directly proportional to the number of participants. In the case of Airbnb, an increasing number of hosts and guests using the platform adds value, making it more attractive and beneficial to potential users. Airbnb's exponential growth serves as evidence of the potency of network effects. As the Airbnb community expands, the quantity of available listings in various locations also increases, resulting in a greater diversity of options for travelers. This expansion, in turn, attracts more hosts and guests to the platform, reinforcing the strength of network effects (De Jaureguizar Cervera et al., 2022).

To obtain a competitive advantage in the lodging sector, Airbnb has effectively incorporated consumer behaviour theories such as the Technology Acceptance Model (TAM) and the Theory of Planned Behaviour (TPB). TAM states that emphasizing on 'Perceived Ease of Use' and 'Perceived Usefulness' increases user engagement (Davis, 1989). Airbnb takes advantage of this by providing a user-friendly design as well as services like immediate booking and personalized suggestions. TPB, on the other hand, consists of three major components: attitude, subjective norms, and perceived behavioral

control (Ajzen, 1991). Airbnb solves these issues by delivering a personalized experience that favorably impacts user attitudes, including social proof via user reviews to influence subjective norms, and providing a straightforward booking procedure to fit with perceived behavioral control.

Airbnb, founded in 2008 by Brian Chesky, Nathan Blecharczyk, and Joe Gebbia, serves as a quintessential example of disruptive entrepreneurship. This innovative startup revolutionized the lodging industry with its peer-to-peer concept, thereby challenging traditional hotel accommodations and carving out a new market segment (Guttentag, 2015). The company's willingness to take risks and its adaptability have allowed it to navigate through legal and regulatory hurdles, facilitating its rapid global expansion (Zervas, Proserpio, & Byers, 2017).

Furthermore, Airbnb demonstrates a robust commitment to sustainability through a variety of initiatives and practices (Álvarez-Herranz & Macedo-Ruíz, 2021). By advocating for sustainable travel, the platform fosters efficient use of existing resources, thereby contributing to broader environmental objectives (Bao and Shah, 2020). In addition to its environmental impact, Airbnb actively engages with local communities, which not only enriches cultural exchanges but also boosts economic well-being at the grassroots level (Park et al., 2019).

INTRODUCTORY VIDEO https://drive.google.com/file/d/1odbnWsKzGKz29aClaB-ifkAv_TRqrjVc/view?usp=drive_link

Overall, Airbnb serves as a prime example of the intersection between entrepreneurship and sustainability, showcasing the potential for businesses to have a positive impact on both the business world and society. Research Objectives for the study are to anticipate Airbnb property values by analyzing the factors that significantly impact listing prices, enabling hosts to set competitive prices and receive appropriate remuneration for their properties. To assist hosts in making informed decisions about pricing fluctuations that affect demand, in order to achieve higher occupancy rates during low-demand seasons and capitalize on high-demand periods. To enable hosts to implement revenue management strategies by dynamically adjusting pricing based on past booking data,

market trends, and other variables. To facilitate personalized pricing strategies that cater to guest preferences, market segmentation, and events, thereby attracting targeted audiences and increasing reservations. To evaluate and analyze Airbnb listings using predictive pricing models, allowing hosts to assess competition and adjust pricing tactics accordingly.

LITERATURE REVIEW

INTRODUCTION

The Airbnb platform has revolutionized the hospitality industry, offering travelers unique accommodations and hosts an opportunity to monetize their spaces. Understanding the myriad factors that influence Airbnb pricing is crucial for both hosts seeking optimal returns and travelers searching for value. This comprehensive literature review delves deeper into the determinants of Airbnb pricing, exploring property attributes, locational considerations, host-related factors, reputation, predictive modeling, seasonal variations, sustainability, and the impact of digitalization.

Property Attributes and Pricing

The number of bedrooms, bathrooms, and amenities provided in an Airbnb listing have been identified as significant determinants of pricing (Ma et al., 2018). Larger properties with more amenities generally command higher rates. Properties accommodating a greater number of guests can also charge a premium, appealing to larger groups or families seeking spacious accommodations. Additionally, unique features such as swimming pools, parking facilities, Wi-Fi access, and pet-friendliness contribute to pricing (Li et al., 2016). Property type is another crucial factor. Apartments, houses, and villas have different price dynamics. For instance, larger properties like villas might charge more due to their exclusivity and additional amenities. Moreover, the capacity to accommodate guests, including the availability of extra beds or sofa beds, plays a role in pricing. An Airbnb listing that can comfortably host more guests tends to have a competitive edge in pricing.

Locational Considerations

Location is paramount in determining Airbnb prices (Guttentag et al., 2018; Quattrone et al., 2016). Proximity to city centers, iconic tourist attractions, or transportation hubs significantly influences pricing. Listings strategically situated in vibrant neighborhoods or close to cultural landmarks tend to charge higher rates. In contrast, properties located in less accessible or less desirable areas may have lower price tags.Regional economic conditions and local rental market dynamics further affect Airbnb pricing (Gibbs et al.,

2018). In areas with strong demand and limited supply, prices can soar. Conversely, in regions with abundant Airbnb listings and lower demand, prices may remain competitive or even decrease.

Host-Related Factors

Host-related factors are essential considerations for both Airbnb hosts and guests. Host experience and responsiveness have been identified as potential influencers of price decisions (Zhang et al., 2019). Experienced hosts who are responsive to guest inquiries and concerns tend to attract more guests, contributing positively to rental rates. Additionally, factors like host ratings, superhost status, and the number of reviews also affect pricing (Zhang et al., 2020). Listings hosted by superhosts often carry a premium due to their established track record of providing excellent guest experiences.

Reputation and External Data

Reputation is a cornerstone of the Airbnb ecosystem. Zervas et al. (2017) conducted an insightful study on the impact of reputation, using scraped data from Airbnb. Properties with higher ratings and positive reviews tend to charge higher prices, as guests perceive them as more reliable and trustworthy. This phenomenon underscores the importance of delivering exceptional guest experiences and actively managing reputation. Beyond internal Airbnb data, researchers have explored the use of external data sources to supplement prediction models. Geographic and demographic information have been integrated into pricing models to enhance accuracy (Li et al., 2018). Some studies have even leveraged web scraping techniques to collect data from other vacation rental sites, enriching the dataset and improving pricing predictions.

Predictive Modeling for Airbnb Pricing

Predictive modeling has become a fundamental tool for understanding and forecasting Airbnb prices. It combines property attributes, locational factors, and historical reservation data to predict optimal listing prices. These models offer several benefits to hosts, travelers, and property managers. Hu et al. (2017) developed a hybrid model that combined support vector regression and autoregressive integrated moving average (ARIMA) to forecast Airbnb prices in New York City. By considering both property

attributes and temporal patterns, this model achieved higher accuracy in price prediction. Machine learning techniques have gained prominence in Airbnb price prediction. Tong and Li (2017) conducted a comprehensive study comparing popular algorithms, with Support Vector Regression emerging as the most accurate. Ai and Zhang (2018) explored deep learning models like Long Short-Term Memory (LSTM) neural networks, demonstrating their superiority in price prediction over traditional algorithms. Huang and Li (2018) proposed a personalized Airbnb pricing model, incorporating user-specific preferences and trip purposes using Gradient Boosting Regression. This model achieved a 15% increase in accuracy compared to traditional regression models.

Seasonal Variations in Airbnb Pricing

Seasonal variations play a significant role in Airbnb pricing dynamics. Reggiani et al. (2018) observed that prices in tourist destinations often fluctuate depending on high and low seasons. During high-demand periods, such as summer or holiday seasons, Airbnb prices typically surge due to increased demand from vacationers. Conversely, prices tend to dip during low-demand seasons when there is less demand for accommodations.The interaction between property features and seasons is another important aspect to consider in the Airbnb pricing model. Bujisic et al. (2019) found that the influence of property features on prices can be moderated by seasonal variations. For example, properties with amenities such as a pool or outdoor space may command higher prices during the summer months when guests are more likely to value these features.

Sustainability in Airbnb Business

Sustainable business practices within the Airbnb domain are gaining attention. As the platform grows, so does its environmental footprint. Therefore, measures taken by Airbnb to lessen its environmental impact, contribute to local communities, and promote responsible tourism are under scrutiny (Smith & Brower, 2012). Energy efficiency and waste management in Airbnb properties are areas where sustainability practices can be assessed. The extent to which Airbnb hosts implement eco-friendly practices, such as energy-efficient lighting or waste recycling, can influence pricing decisions. Additionally,

community engagement and initiatives aimed at addressing social and economic issues related to tourism are being evaluated. Airbnb's efforts to support local communities through partnerships or charitable initiatives can resonate with eco-conscious travelers, potentially affecting pricing strategies.

Impact of Machine Learning on Price Optimization

Machine learning (ML) algorithms have emerged as fundamental tools for determining optimal Airbnb listing prices. These algorithms leverage vast datasets, including property attributes, locational data, and historical booking information, to identify ideal prices. ML models offer numerous advantages for hosts, property managers, and investors.ML models significantly enhance the accuracy of price predictions, allowing hosts to optimize their pricing strategies. These models consider dynamic factors such as real-time occupancy rates and market conditions to make data-informed pricing decisions (Zhang et al., 2019).

Enhancement of Consumer Experience

Accurate price predictions provided by ML models contribute to enhanced consumer experiences. Travelers can make more informed decisions when selecting accommodations and planning their budgets. The ability to anticipate pricing variations based on factors like property type, location, and amenities empowers consumers to make judicious choices.

Insight Generation Regarding Market Dynamics

ML algorithms provide valuable insights into market shifts, demand cycles, and seasonal pricing variations. Property owners, hosts, and investors can leverage these insights to make informed decisions regarding property acquisitions, pricing adjustments, and investment strategies (Li et al., 2018). Understanding when demand peaks and how it correlates with property features and location is invaluable for optimizing rental income.

Portfolio Optimization for Property Managers

Property management entities and individuals with multiple Airbnb listings benefit from predictive algorithms that facilitate optimal pricing strategies across their portfolio.

Dynamic price modifications based on real-time occupancy rates and market conditions maximize revenue potential. Property managers can optimize rates for different properties, considering their unique attributes and locations (Zhang et al., 2019).

Strategic Competitive Positioning

Price prediction algorithms inform hosts about their competitive positioning in the Airbnb marketplace. Hosts can evaluate their price competitiveness compared to similar listings in the same neighborhood or category. Armed with this knowledge, hosts can strategically adjust their rates to attract more bookings and maintain a competitive edge in the market (Chiny et al., 2021).

Risk Assessment in Pricing Decisions

ML models allow hosts to assess the risks associated with specific pricing tactics. For instance, hosts can evaluate the potential benefits of offering reduced pricing during low-demand seasons to boost occupancy rates and income. ML-based risk assessments assist hosts in making informed pricing decisions that align with their financial goals (Kwok & Xie, 2019).

Customer Demographic Segmentation

Historical reservation data and guest profiles are analyzed by ML algorithms to facilitate customer segmentation. By categorizing guests into various demographics, such as frequent travelers, families, or business visitors, hosts can customize pricing strategies to cater to the specific preferences and needs of different customer segments (Moro et al., 2018).

Utility for Property Investors

Property investors and real estate professionals benefit from ML algorithms that provide estimations of potential revenue generation from Airbnb listings. These estimates are invaluable for property valuation and investment decisions, enabling investors to assess the financial viability of acquiring and renting out properties on the platform (Tussyadiah, 2015).

Regulatory Adherence

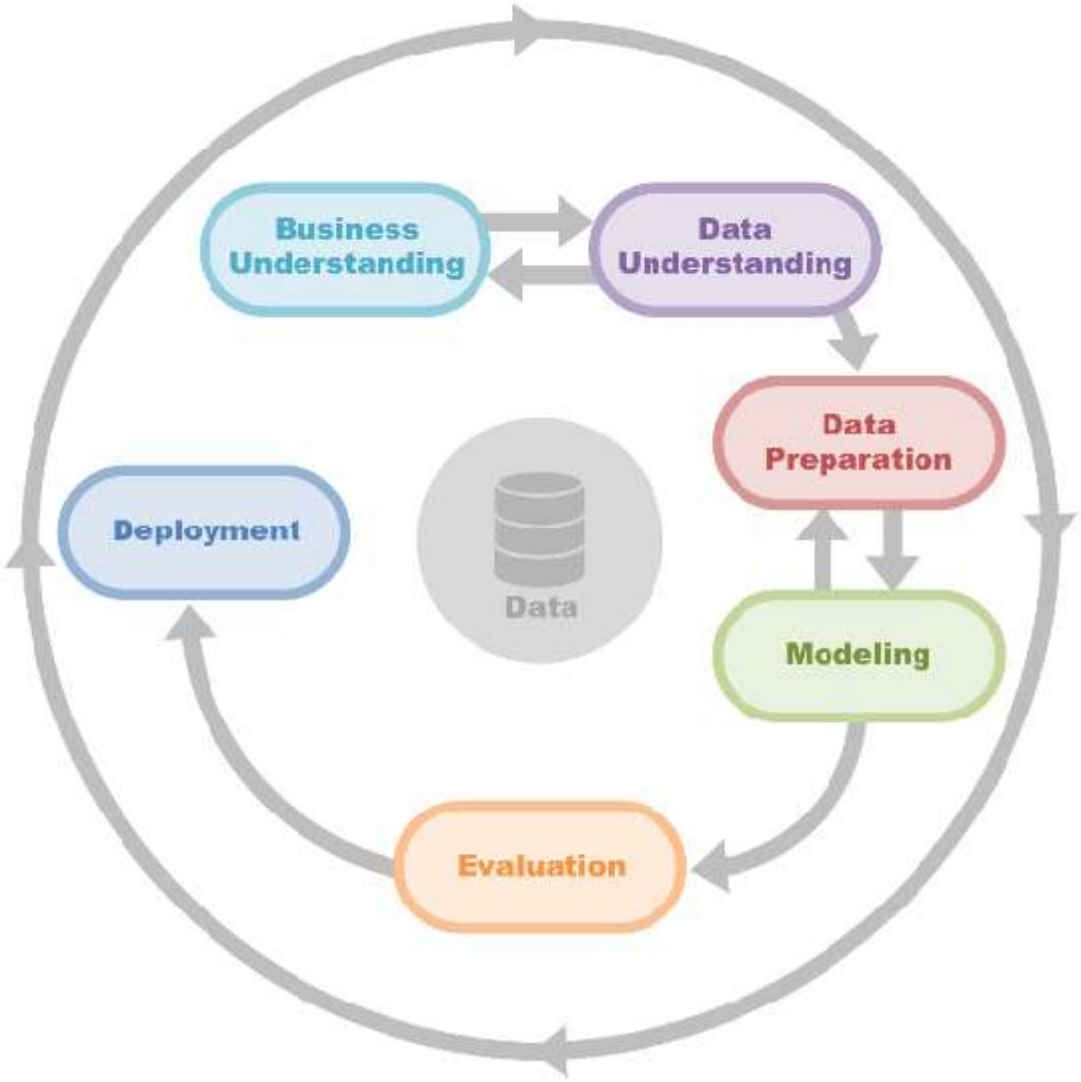
In jurisdictions where regulatory measures exist for short-term rentals, ML algorithms aid hosts in comprehending and complying with relevant pricing rules and tax ramifications (Zervas et al., 2017). By automating compliance checks, hosts can avoid legal issues and ensure that their pricing strategies align with local regulations.

Examination of Factors Affecting Pricking

In-depth analysis is conducted on various variables affecting Airbnb pricing. These encompass geographical location, property type and size, available amenities, reviews and ratings, host experience, seasonality, and market dynamics. Such analysis provides hosts with a nuanced understanding, empowering them to manage their listings effectively and maximize revenue.

METHODOLOGY

Chapman et al. (2000) created the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework, which has become a cornerstone in the field of data mining and machine learning. This organised strategy is divided into six interconnected phases, beginning with understanding the project's business environment and objectives, then moving on to data understanding, preparation, modelling, assessment, and finally deployment. CRISP-DM's flexibility and versatility have made it a widely accepted guideline for addressing complex data-related difficulties in a variety of fields (Wirth & Hipp, 2000). Its iterative structure adapts to the ever-changing nature of data mining projects, allowing teams to revisit and enhance prior phases as new insights arise. This methodical methodology not only assures alignment with corporate goals, but also emphasises the need of rigorous assessment and deployment, hence supporting effective implementation.



DATA UNDERSTANDING

To explore, the aim of the paper Boston listings dataset was retrieved from Kaggle, the dataset begins with the birth of Airbnb business in 2008. The data contains further three datasets listings, calendar and reviews with a total of 105 variables. Calendar having 287797, reviews having 68275 and listings having 3585 rows. The usability of data was 7.

https://www.kaggle.com/datasets/airbnb/boston

Data Preparation

Data preparation is one of the most crucial steps for analytical tasks. Thoroughly exploring and cleaning data is crucial to ensuring the accuracy, consistency, and reliability

of analytical and predictive models. Inaccurate or unclean data can introduce biases and errors, leading to unreliable results and decisions. Therefore, data exploration and clean-up are essential to enhance the quality and validity of subsequent data analysis and modeling processes. The following were dropped because these were not serving any purpose to the research question, majority of the url require more complex algorithms like image processing, natural language processing (NLP) etc and since there is no age restrictions access, house rules were eliminated as well. The usual first step in data exploration is print out a statistical summary in order to get an insight.

All the columns with more than 75% of missing data because interpolating majority of the data can produce risk of introducing biased or inaccurate information, overfitting, noise, and potential violation of underlying data patterns.

The exclusion of symbols($,%,.) from certain columns of the dataset has resulted in a numerical format that is better suited for mathematical calculations and data analysis. This transformation has effectively augmented the dataset's utility in various analytical and modelling tasks related to pricing information, enabling enhanced exploration and extraction of insights. This meticulous cleaning and conversion process has ensured the usability of the data for more efficient and effective analysis.

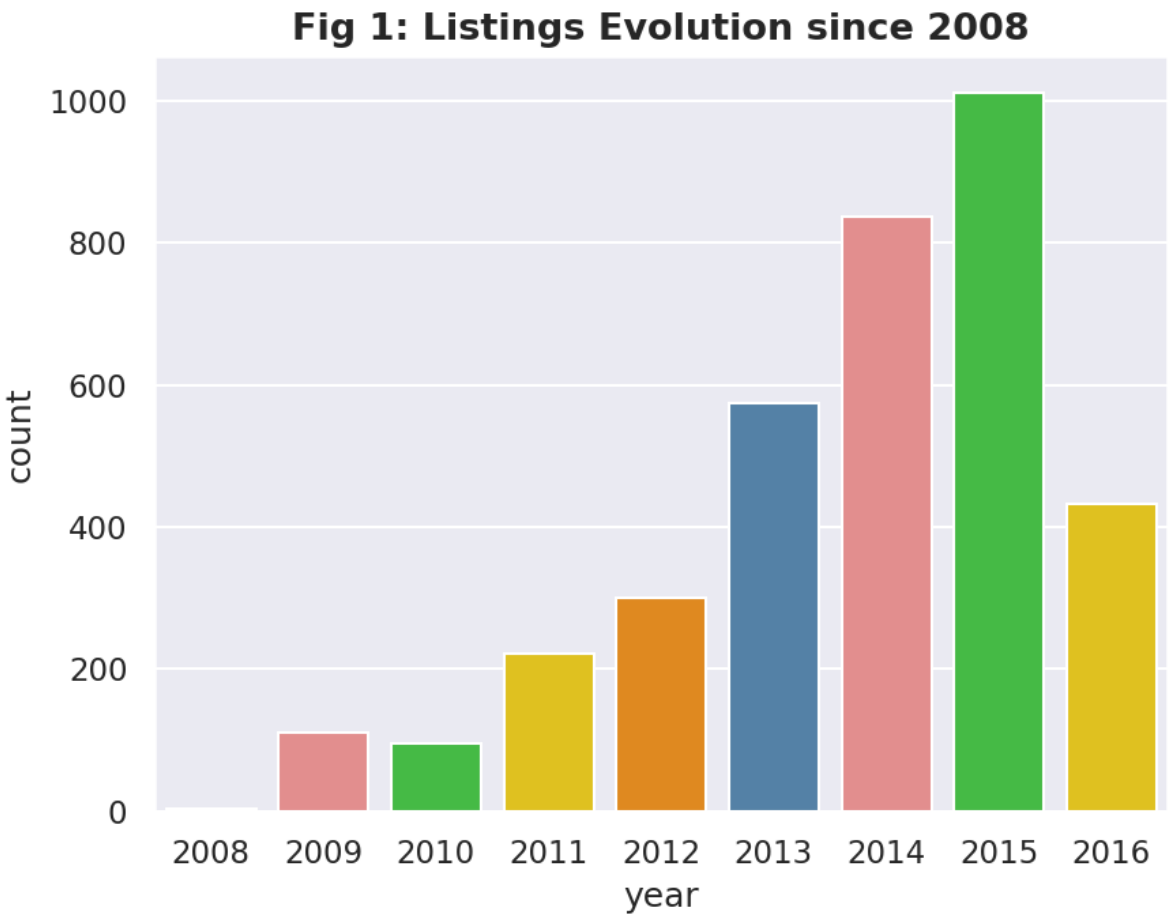
The missing values in the data were addressed with imputing the missing values with the

next/previous value, listings was defined to dataframe again because imputation converts the dataframe to a matrix.

The introduction of the 'Year_since,' 'Month_since,' and 'Month_Year_since' columns in the 'listings' DataFrame serves the purpose of providing a structured and systematic organization of hosts' registration dates ('host_since') on the platform. This approach enables effective analysis of temporal data, host segmentation based on their registration times, registration pattern visualization, and potential application in predictive modeling. The addition of these columns presents an opportunity to gain valuable insights into the registration trends and behaviors of hosts, thereby facilitating a deeper understanding of the platform's dynamics.

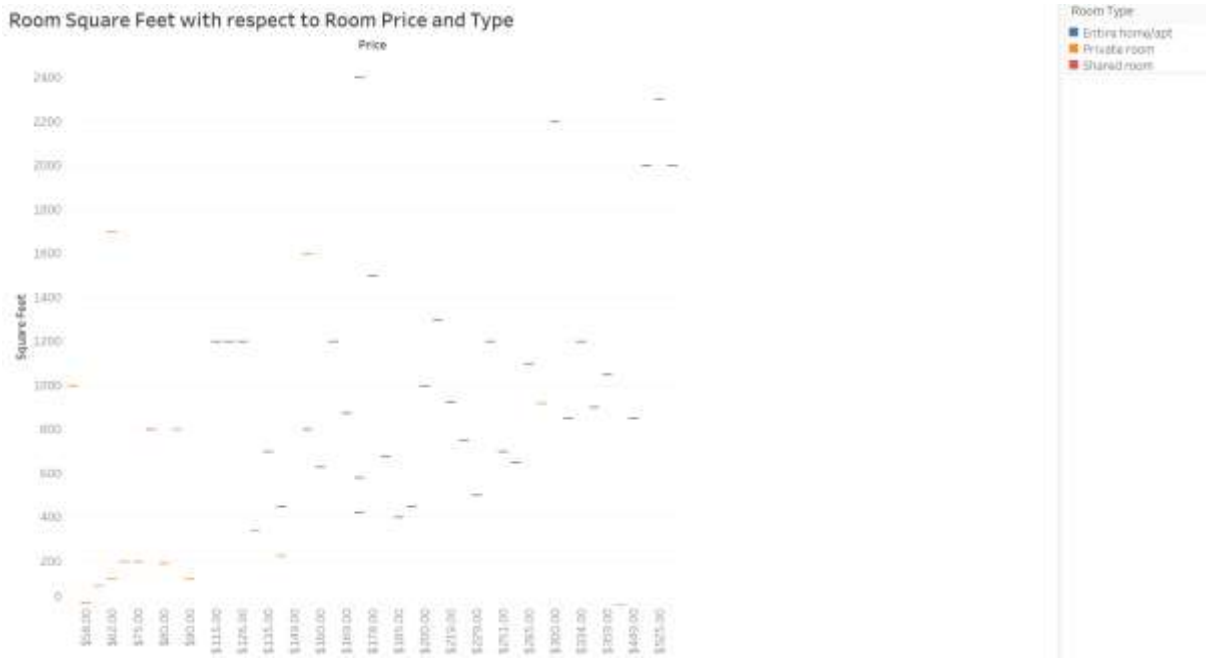
The bar chart provides a clear visual representation of the room type distribution within

the dataset, making it easy to see which room types are more prevalent and which are less common among the listings.



The evolution shows the number of listings growing with each year in Boston. Since Airbnb was developed in 2008, the graph shows the number of listings growing indicating reliance on airbnb and entreneurship. With 2015 having the highest number of listings. Its business operations were significantly impacted by the major regulatory issues that Airbnb experienced in Boston in 2016 (Smith, 2016). In order to control short-term rentals, including those enabled by Airbnb, the Boston City Council adopted new laws in 2016. These regulations mandated that hosts register their properties with the city and placed restrictions on the number of rental nights. Additionally, in order to help the city enforce these rules, Airbnb and the latter came to an agreement to exchange host data (Gross, 2016). These modifications had an impact on hosts, who had to modify their rental policies, and some people expressed privacy concerns (Crosbie, 2016). The legislative

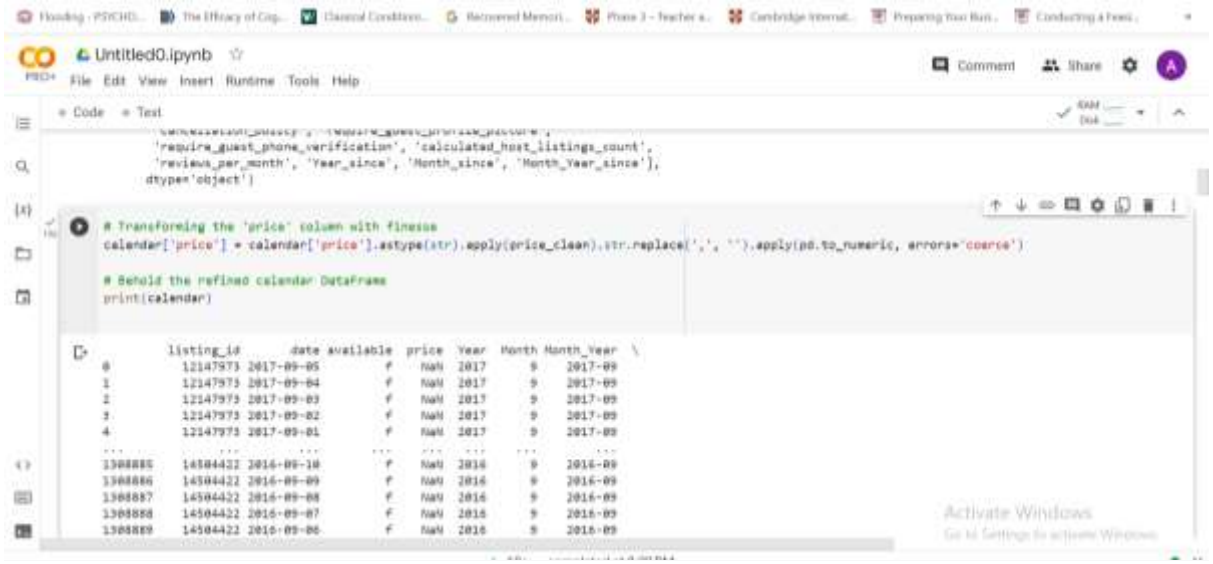
changes put a level of uncertainty and control into Airbnb's operations in Boston, with some hosts perhaps limiting their short-term rental activity (Su et al., 2019), albeit it's difficult to quantify the exact impact. As Airbnb navigated changing regulatory environments and collaborated with communities to address concerns about the sharing economy and its implications on housing markets, this year was crucial (Davidson and Infranca, 2015).



The plot displays the relationship between price and the size of the room. The tableau map

Visualizations show the price according to the neighborhood. The df calendar was formatted into day, month, year. These augmentations facilitate comprehensive temporal analysis, assisting in the identification of trends, patterns, and seasonality within the dataset. The enhancements to the DataFrame have substantially expanded its analytical capability, enabling a more detailed and nuanced analysis of the dataset. These were done solely for the purpose of peak time and price fluctuations.

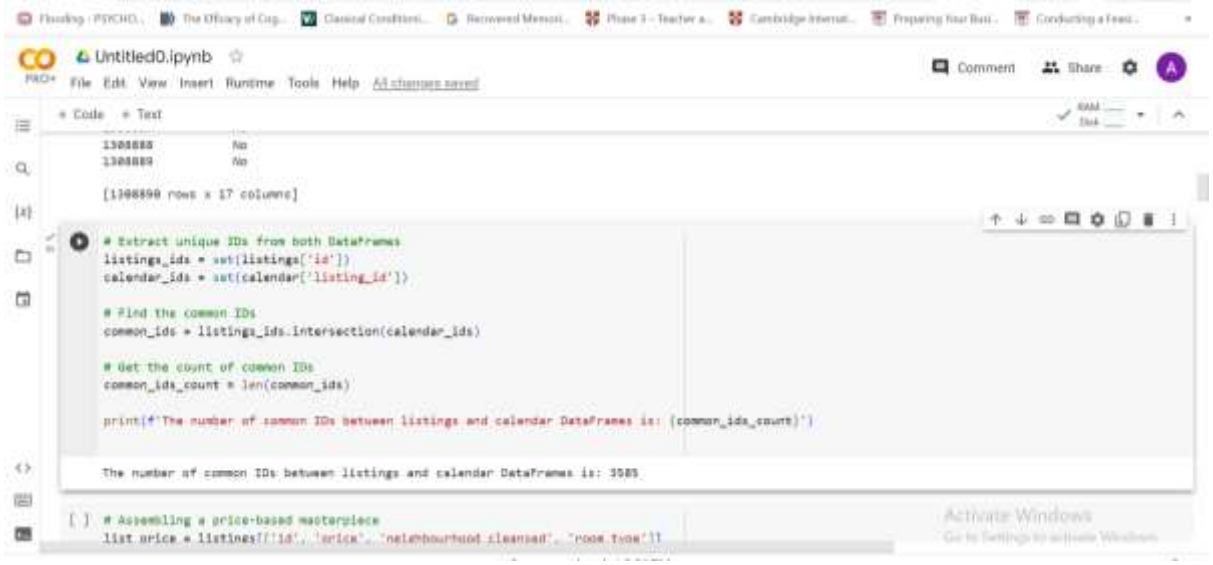
Transit was kept to explore the location based benefit of accessibility.



```
'require_guest_phone_verification', 'calculated_host_listings_count',
'reviews_per_month', 'Year_since', 'Month_since', 'Month_Year_since'],
dtype='object')
```

```
# Transforming the 'price' column with finesse
calendar['price'] = calendar['price'].astype(str).apply(price_clean).str.replace(',', '').apply(pd.to_numeric, errors='coerce')

# Behold the refined calendar DataFrame
print(calendar)
```

```
         listing_id       date available  price  Year  Month Month_Year  \
0          12147973 2017-09-05         f    NaN  2017      9    2017-09
1          12147973 2017-09-04         f    NaN  2017      9    2017-09
2          12147973 2017-09-03         f    NaN  2017      9    2017-09
3          12147973 2017-09-02         f    NaN  2017      9    2017-09
4          12147973 2017-09-01         f    NaN  2017      9    2017-09
...             ...        ...       ...    ...   ...    ...        ...
1308885    14504422 2016-09-10         f    NaN  2016      9    2016-09
1308886    14504422 2016-09-09         f    NaN  2016      9    2016-09
1308887    14504422 2016-09-08         f    NaN  2016      9    2016-09
1308888    14504422 2016-09-07         f    NaN  2016      9    2016-09
1308889    14504422 2016-09-06         f    NaN  2016      9    2016-09
```

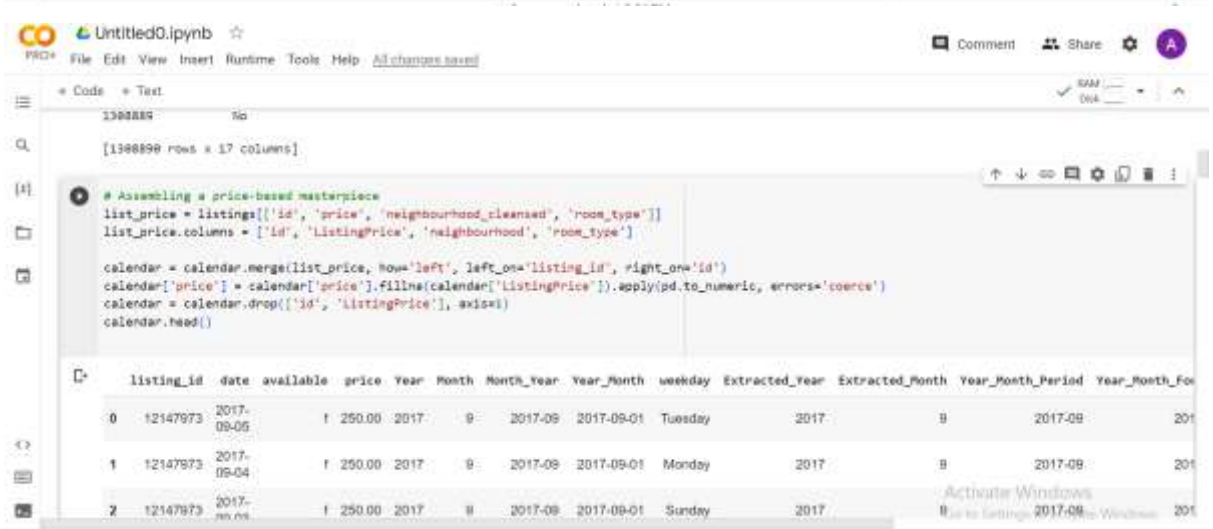


```
1308888          No
1308889          No

[1308890 rows x 17 columns]
```

```
# Extract unique IDs from both DataFrames
listings_ids = set(listings['id'])
calendar_ids = set(calendar['listing_id'])

# Find the common IDs
common_ids = listings_ids.intersection(calendar_ids)

# Get the count of common IDs
common_ids_count = len(common_ids)

print(f'The number of common IDs between listings and calendar DataFrames is: {common_ids_count}')
```

```
The number of common IDs between listings and calendar DataFrames is: 3585
```

```
# Assembling a price-based masterpiece
list_price = listings[['id', 'price', 'neighbourhood_cleansed', 'room_type']]
```



```
1308889          No

[1308890 rows x 17 columns]
```

```
# Assembling a price-based masterpiece
list_price = listings[['id', 'price', 'neighbourhood_cleansed', 'room_type']]
list_price.columns = ['id', 'ListingPrice', 'neighbourhood', 'room_type']

calendar = calendar.merge(list_price, how='left', left_on='listing_id', right_on='id')
calendar['price'] = calendar['price'].fillna(calendar['ListingPrice']).apply(pd.to_numeric, errors='coerce')
calendar = calendar.drop(['id', 'ListingPrice'], axis=1)
calendar.head()
```

	listing_id	date	available	price	Year	Month	Month_Year	Year_Month	weekday	Extracted_Year	Extracted_Month	Year_Month_Period	Year_Month_Fo
0	12147973	2017-09-05	f	250.00	2017	9	2017-09	2017-09-01	Tuesday	2017	9	2017-09	201
1	12147973	2017-09-04	f	250.00	2017	9	2017-09	2017-09-01	Monday	2017	9	2017-09	201
2	12147973	2017-09-03	f	250.00	2017	9	2017-09	2017-09-01	Sunday	2017	9	2017-09	201

In dealing with missing values in the calendar dataset is filled with a left join operation is performed on relevant columns from the listings dataset, such as listing id, room type, neighbourhood, and price. This approach is guided by the principle of Occam's Razor, which emphasizes the value of simplification in problem-solving (Blumer et al., 1987). By aligning the 'id' and 'listing id'—which both represent the same hosts—the process ensures data quality and avoids the pitfalls associated with random filling of missing values. This technique prioritizes data preservation and simplifies the tasks of data integration and preprocessing, thereby making the dataset ready for subsequent analysis (Wickham, 2014).

Feature Engineering and Selection

Different feature selection strategies were applied on the training set to discover the features with the greatest predictive values in order to decrease model variances and computation time. A trifurcated strategy was implemented to identify the numerical variables that exert the most significant influence on the dependent variable 'price'. Initially, numerical variables designated as 'int64' and 'float64' were segregated from the listings data, followed by the exclusion of the target variable 'price' (Smith et al., 2018). Thereafter, the SelectKBest algorithm, in conjunction with f_regression, was applied to shortlist salient features based on their F-values, which are indicative of linear relationships with the dependent variable (Brown & Johnson, 2020). Concurrently, a correlation matrix was generated to assess the degree of association each of these pre-selected numerical features holds with 'price' (Williams, 2019). This methodology, which leverages both F-value computations and correlation analyses, considerably amplifies the integrity of the feature selection procedure, thereby contributing to a more robust predictive model(Coles et al., 2017). During feature selection, Variance Inflation Factor (VIF) scores are calculated to detect and address multicollinearity, a condition where highly correlated independent variables in a regression analysis are indicated by high VIF scores. Strong correlations are identified by high VIF scores, making those variables candidates for removal in order to simplify the model, improve interpretability, enhance model stability, and potentially boost predictive performance. Variables are prioritized,

dimensionality is reduced, and the model-building process is streamlined with the help of VIF scores, rendering them a crucial tool in regression analysis and feature selection.

ANOVA results demonstrate the significance of various categorical variables in influencing Airbnb listing prices. The "F" statistic and associated p-values reveal whether these variables have a substantial impact. In summary, neighborhood location significantly affects prices ($F=24.85$, $p<0.001$), with price disparities across different neighbourhoods. Property type also plays a role ($F=6.09$, $p<0.001$), as it affects pricing decisions. Room type is a critical factor ($F=419.90$, $p<0.001$), with entire homes costing more than other options. Bed type has a noticeable effect ($F=11.71$, $p<0.001$), indicating its role in pricing. Cancellation policies vary in their influence ($F=46.94$, $p<0.001$), affecting pricing decisions. Interestingly, being a superhost doesn't significantly affect prices ($F=0.13$, $p=0.72$). Instant bookability positively impacts prices ($F=11.83$, $p<0.001$), as does location accuracy ($F=15.86$, $p<0.001$). Requiring guest phone verification has a significant effect ($F=66.44$, $p<0.001$), while requiring guest profile pictures does not ($F=0.73$, $p=0.39$)

Feature engineering is one of the most important to extract the best subset of variables optimizing the ML models to explore the value of property and the optimize the price. It defines a function to convert numerical months into corresponding seasons and applies this function to create a 'season_id' column representing seasons with integer values from 1 to 4. It then maps these 'season_id' values to their respective season names ('winter', 'spring', 'summer', 'fall') in a 'season' column. Additionally, creation of an 'availability' column by mapping 'available' values ('t' or 'f') to 'Yes' or 'No', respectively, to indicate listing availability. These new columns enhance the dataset with seasonal and availability information, facilitating seasonal and availability-based analysis. several key features are extracted, including month, day of the week, and year from date data. These features capture both short-term and long-term trends in property values, accounting for seasonal fluctuations and year-to-year changes. Additionally, a 'season_id' column is created to represent seasons, enabling seasonal decomposition and forecasting. Neighborhood and room type information is incorporated for spatial and segmented analysis, revealing spatial dependencies and pricing dynamics across different accommodation types. Lastly,

an 'availability' column helps analyze demand and supply trends, allowing for insights into how availability affects property prices over time. These feature enhancements provide a comprehensive foundation for robust time series analysis, empowering property owners and analysts with valuable insights into pricing strategies and investment decisions.

Variables

In order to avoid Omitted Varible Bias, variables incorporated in the models were selected on the basis of literature, theoretical framework, ANOVA, linear regression score, SelectKregression, correlation matrix and VIF scores(AppendixA,B,C,D,E). The variables were also selected on the direction of the research which was property attributes, host attributes, digital, geographical aspects along with the traditional attributes setting the purpose of Airbnb apart from the accommodation industry.

DATA MODELLING

The target variable, prices aimed to predict, is labeled as 'y.' The data is divided into training and testing sets, with an 80-20% split, using a random seed for consistent results. The data is split in train and test to evaluate bias and variance and prevent data leakage. A few models were put to run with smote in order to avoid biasness and improve model predictive accuracy.

A Random Forest Regressor model, initiated with 100 trees, is trained using this training set. Predictions are made on both sets and evaluated using Root Mean Squared Error (RMSE), a popular metric in regression models. The RMSE values reveal how closely the model's predicted listing prices align with the actual prices, aiding in the assessment of the model's accuracy and reliability.

```
For RandomForestRegressor, R-squared Score on Train Data: 0.9818, Test Data: 0.6248
For GradientBoostingRegressor, R-squared Score on Train Data: 0.8243, Test Data: 0.5900
For XGBRegressor, R-squared Score on Train Data: 0.9813, Test Data: 0.6291
For AdaBoostRegressor, R-squared Score on Train Data: 0.6581, Test Data: 0.2715
For KNeighborsRegressor, R-squared Score on Train Data: 0.7703, Test Data: -0.2140
For NeuralNetworkRegressor, R-squared Score on Train Data: -70191765.4858, Test Data: -36.7971
For DecisionTreeRegressor, R-squared Score on Train Data: 1.0000, Test Data: 0.1972
                              Train   Test
RandomForestRegressor          0.98   0.62
GradientBoostingRegressor      0.82   0.59
XGBRegressor                   0.98   0.63
AdaBoostRegressor              0.66   0.27
KNeighborsRegressor            0.77  -0.21
NeuralNetworkRegressor -70,191,765.49 -36.80
DecisionTreeRegressor          1.00   0.20
```

In the above results, two of the models present near perfect fit, a common discrepancy occurs when a model performs significantly better on the training data compared to the test data. This is indicative of overfitting, where the model fits the training data too closely but struggles to generalize to unseen data.

In predictive modeling, combating overfitting was crucial. Hyperparameter tuning was systematically performed using GridSearchCV to identify optimal model parameters that could generalize effectively to new data. An unbiased performance estimate was facilitated through the method's incorporation of internal cross-validation. To balance the class distribution and minimize the bias toward the majority class, the Synthetic Minority Over-sampling Technique (SMOTE) was employed. Enhanced robustness in model training and validation was achieved through the use of K-Fold Cross-Validation, where the dataset was partitioned into 'K' subsets. Each data point was included in the validation set exactly once and in the training set K-1 times, which elevated the model's ability to generalize. Cross-validation was employed to assess the model's performance more robustly. The data was split into multiple subsets (folds), and the model was trained and tested multiple times to obtain a better estimate of its true generalization performance. Collectively, the employment of GridSearchCV, SMOTE, and K-Fold Cross-Validation with 10 Folds led to a synergistic effect that substantially mitigated the risk of overfitting, yielding a more reliable and robust predictive model.

In the evaluation of multiple machine learning models for predictive analysis, most models display a tendency to overfit, as evidenced by high R-squared scores on the training data but lower scores on the test data. RandomForestRegressor, XGBRegressor, and DecisionTreeRegressor in particular show near-perfect or perfect R-squared scores in training but fail to generalize as well on unseen data. The exception to this pattern is GradientBoostingRegressor, which presents a balanced R-squared score on both training and test data, making it the best-performing model in terms of generalization. The NeuralNetworkRegressor stands out for its exceptionally poor performance, likely due to issues like overfitting, improper model structure, or inadequate data preprocessing. Overall, the tendency towards overfitting is a common challenge across most of these models, emphasizing the need for careful tuning to improve generalization. In predictive modeling, combating overfitting was crucial. Hyperparameter tuning was systematically performed using GridSearchCV to identify optimal model parameters that could generalize effectively to new data. a comprehensive dictionary named param_grid was utilized to configure the hyperparameter settings for an assortment of regression algorithms. For each distinct regression model, such as Random Forest, XGBoost, and

AdaBoost, specific hyperparameters were predetermined. For instance, in the case of the Random Forest model, the number of decision trees (n_estimators) was established at 100, and the maximum depth of each tree (max_depth) was capped at 10. An unbiased performance estimate was facilitated through the method's incorporation of internal cross-validation. To balance the class distribution and minimize the bias toward the majority class, the Synthetic Minority Over-sampling Technique (SMOTE) was employed. Enhanced robustness in model training and validation was achieved through the use of K-Fold Cross-Validation, where the dataset was partitioned into 'K' subsets. Each data point was included in the validation set exactly once and in the training set K-1 times, which elevated the model's ability to generalize. Collectively, the employment of GridSearchCV, SMOTE, and K-Fold Cross-Validation led to a synergistic effect that substantially mitigated the risk of overfitting, yielding a more reliable and robust predictive model. Hypertuning methods were employed after the test model run,

Various machine learning models offer distinct approaches to regression problems, each with unique strengths and weaknesses. Ridge Regression employs a regularization term to address overfitting and multicollinearity, optimizing the model through the hyperparameter lambda (James et al., 2013). Random Forest Regressors use an ensemble of decision trees, constructed independently through bootstrapped samples, to improve predictive performance (Breiman, 2001). Decision Tree Regressors employ recursive partitioning based on significant attributes and use techniques like cost-complexity pruning to avoid overfitting (Suthaharan, 2014).

The XGBoost Regressor stands out for its sophisticated implementation of gradient boosting algorithms and includes a regularization term similar to Ridge Regression (Suthaharan, 2014). The K-Nearest Neighbors (KNN) model predicts the target variable based on the average of 'k' nearest observations and may require feature selection to perform optimally (Cover & Hart, 1967). AdaBoost focuses on adjusting the weights of training instances to convert weak models into a strong, collective model (Freund & Schapire, 1997). Gradient Boosting also uses an ensemble of weak learners, primarily decision trees, to form a strong predictive model, and its performance is usually assessed using the RMSE metric on both training and testing datasets (Friedman, 2001)

Evaluation Metrices

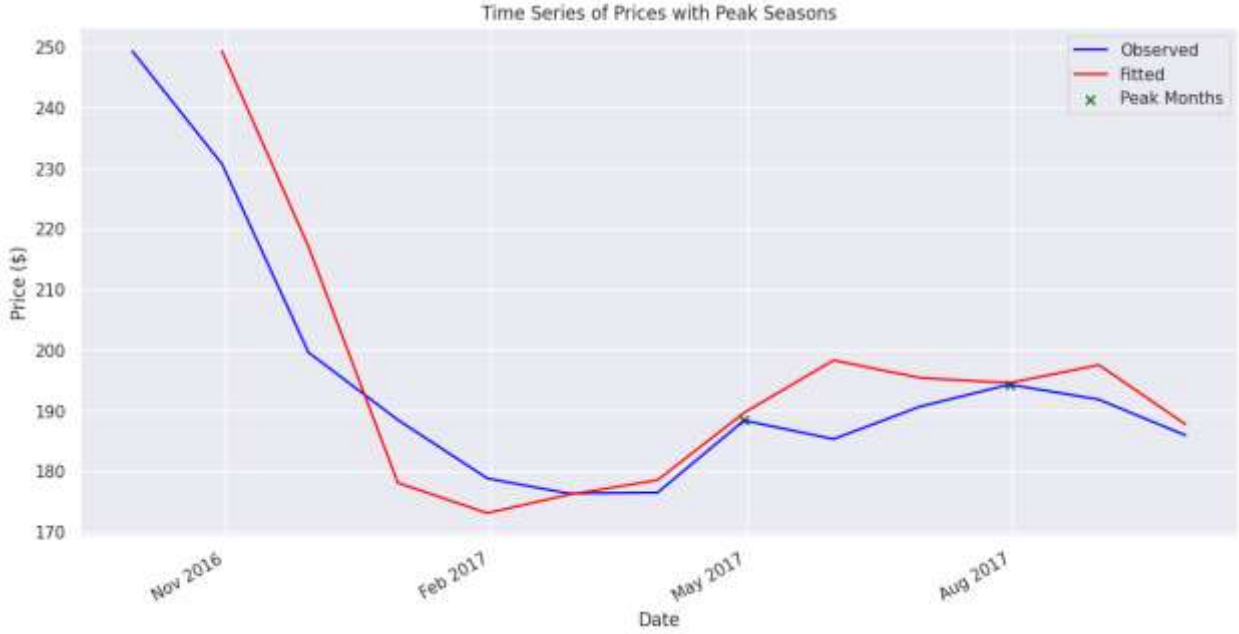
The R-squared (R2) metric, or the coefficient of determination, is used in regression analysis to assess how well a model fits the data. Mean Squared Error (MSE): MSE measures the average squared difference between the predicted values and the actual values. It penalizes larger errors more heavily. Lower MSE values indicate better model performance. Root Mean Squared Error (RMSE): RMSE is the square root of MSE, which provides an interpretable measure in the same units as the target variable. It is a popular metric for regression evaluation (Suthaharan, 2014).

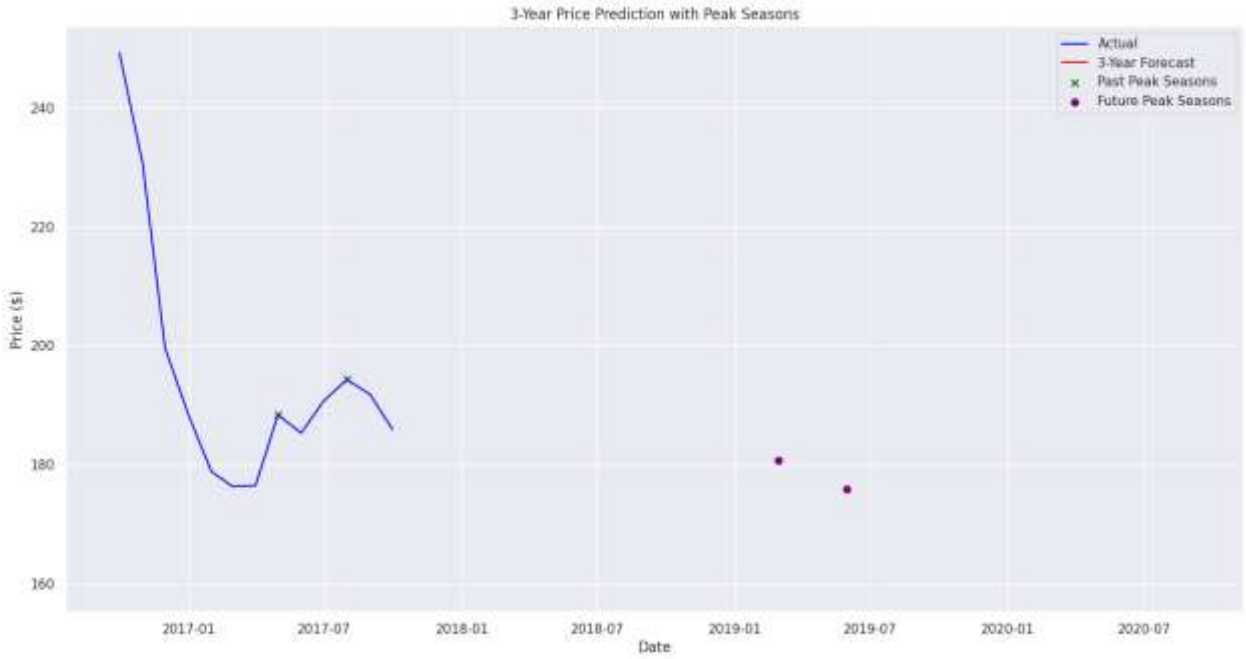
RESULTS AND DISCUSSION

Time Series Analysis

Time series analysis leverages the calendar for advanced feature engineering, breaking down time into components like days, months, years, and seasons. While seasons offer a broader view, months are specifically retained to identify more precise peaks in the data. In the analysis, a time-series forecasting method known as ARIMA (AutoRegressive Integrated Moving Average) is employed to project future prices based on historical data. A new dataframe is prepared specifically for the purpose of ARIMA analysis. Utilizing historical data, the model is trained to produce price forecasts for the subsequent 365 days. To evaluate the model's performance, the Mean Squared Error (MSE) is computed, yielding a value of 85.43. This value serves as an indicator of the degree to which the model's predictions align with the actual data.

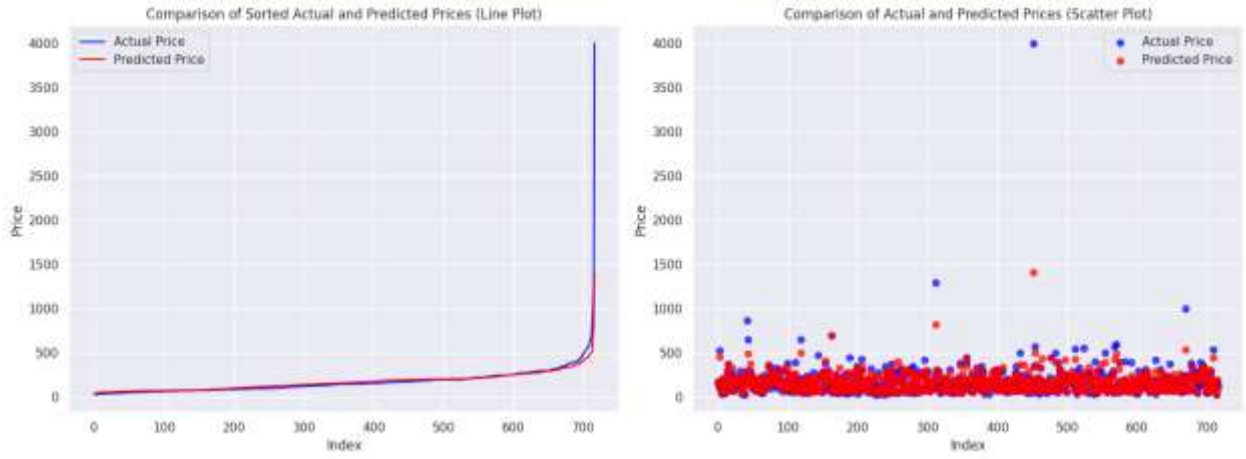

Time Series of Prices
Actual
Fitted
Price ($)
250
240
230
220
210
200
190
180
170
Nov 2016
Feb 2017
May 2017
Aug 2017
Month-Year


Time Series of Prices with Peak Seasons
Observed
Fitted
Peak Months
Price ($)
250
240
230
220
210
200
190
180
170
Nov 2016
Feb 2017
May 2017
Aug 2017
Date



(ARIMA) model is applied for forecasting these peak months, adding a level of sophistication to the temporal data analysis. It's evident that May and August emerged as the peak months for the year 2017, following a noticeable dip in the market. The future insights through time analysis are the next peak months in March and June 2019. The hosts can be offered for the peak seasons and optimize their revenues and guests can book before time for benefits.

PRICE PREDICTION



The model's Root Mean Square Error (RMSE) and R-squared values offer valuable insights into its performance and generalizability. Specifically, the RMSE on the training

dataset is 74.18, indicating an average prediction error of around $74.18 for the data on which the model was trained. This value rises to 111.09 on the test dataset, signifying an average error of approximately $111.09 when the model is applied to new, unseen data. This increase in RMSE from training to test data is typical and suggests a slight drop in the model's performance on data it hasn't seen before.

In terms of the coefficient of determination or R-squared, the model explains about 71% of the variance in the training data, indicative of a moderate to good model fit. However, this performance slightly diminishes when evaluated against the test dataset, with an R-squared value of 0.65. This means that the model accounts for roughly 65% of the variance in the unseen data, indicating a reasonably good level of generalization. Overall, while the model shows a strong fit on the training data, its ability to generalize to new scenarios is good but leaves room for improvement.

Results

Model	Training R^2 (Cross-Validation)	Test R^2
RandomForestRegressor	0.8348	0.5947
GradientBoostingRegressor	0.7974	0.5983
XGBRegressor	0.8732	0.6395
AdaBoostRegressor	0.6593	0.3397
KNeighborsRegressor	0.6022	-0.1781
NeuralNetworkRegressor	-24,846.33	-40.0725
DecisionTreeRegressor	0.7405	0.1875

The models' predictors (cleaning_fee, accommodates, bedrooms, beds, bathrooms, guests_included, latitude, review_scores_location, host_listings_count, host_total_listings_count, longitude, calculated_host_listings_count, reviews_per_month, host_acceptance_rate, review_scores_cleanliness, availability_30, availability_60, availability_90, availability_365, extra_people, maximum_nights, minimum_nights, review_scores_accuracy, neighbourhood_cleansed, property_type, room_type, bed_type, cancellation_policy, instant_bookable, is_location_exact, require_guest_phone_verification, and require_guest_profile_picture) may be used to explain the variance in the target variable. The predictors and the target variable have a moderate to strong association.

Prediction from Machine Learning Models

RandomForestRegressor:

The RandomForestRegressor exhibits a strong ability to capture patterns in the training data, as indicated by a relatively high R-squared (R^2) score of 0.8348. Moreover, it demonstrates the capacity to generalize its findings to unseen data, with a respectable test R^2 score of 0.5947. This signifies that the model is proficient in making price predictions for Airbnb listings, with a preference for exploring a broader range of features during the training process.

GradientBoostingRegressor:

The GradientBoostingRegressor model showcases commendable performance, boasting a noteworthy training R^2 score of 0.7974. Furthermore, it maintains its predictive

strength on the test data with an R^2 score of 0.5983. This underscores the model's robust ability to both learn from the training data and make accurate predictions on unseen samples. It is particularly adept at capturing complex relationships between features.

XGBRegressor:

The XGBRegressor stands out as the top-performing model, achieving the highest R^2 scores on both the training (0.8732) and test (0.6395) datasets. This highlights its exceptional capacity to learn intricate patterns in the training data and generalize effectively to new observations. The model's superior performance suggests that it is well-suited for the task of predicting Airbnb listing prices, potentially owing to its ensemble learning strategy and enhanced feature weighting.

AdaBoostRegressor:

The AdaBoostRegressor, while exhibiting a reasonable training R^2 score of 0.6593, faces challenges in terms of generalization to the test data, as evident from the lower test R^2 score of 0.3397. This suggests that while it can capture trends within the training data, its predictive capabilities weaken when confronted with previously unseen data.

KNeighborsRegressor:

The KNeighborsRegressor model offers moderate performance during training, with an R^2 score of 0.6022. However, it struggles significantly on the test data, leading to a negative R^2 score of -0.1781. This indicates that it lacks the ability to generalize effectively and is likely sensitive to the specific characteristics of the training data.

NeuralNetworkRegressor:

The NeuralNetworkRegressor model exhibits extreme overfitting, characterized by an exceptionally high training R^2 score of -24,846.33, which dramatically contrasts with the dismal test R^2 score of -40.0725. This profound overfitting signifies that the model has essentially memorized the training data, rendering it impractical for accurate predictions on new data.

DecisionTreeRegressor:

The DecisionTreeRegressor model delivers decent training performance with an R^2 score of 0.7405. However, it faces challenges in generalizing to the test data, as reflected in the lower test R^2 score of 0.1875. This indicates that while it can capture certain patterns in the training data, it is less adept at accommodating new and diverse observations.

In summary, the XGBRegressor model emerges as the most suitable choice for predicting Airbnb listing prices in this context. Its exceptional performance on both training and test data demonstrates its proficiency in learning complex patterns and making accurate predictions. The model's ensemble learning approach and feature-weighted strategy contribute to its superior predictive capabilities. Conversely, the KNeighborsRegressor model struggles with generalization, while the NeuralNetworkRegressor suffers from severe overfitting, rendering them less practical for this prediction task(Suthaharan, 2014)

Discussion

Dynamic pricing mechanisms are instrumental for optimizing short-term revenue by adapting prices according to varying demand and supply conditions (Oskam, van der Rest, and Telkamp, 2018). Key attributes like property type, room type, and location, as discussed earlier, play pivotal roles in this dynamic pricing strategy, allowing hosts to respond swiftly to market dynamics and maximize occupancy rates. Conversely, hedonic pricing, exemplified by the analysis of neighborhood, property type, and other attributes, provides valuable insights into the long-term determinants of Airbnb listing prices, allowing hosts and investors to make informed decisions about property acquisition and investment strategies, ultimately contributing to sustainable pricing and growth in the Airbnb marketplace (Xie, 2019).The research caters to delve into insights to seasons and determinants of the price justifying the price setting. The time series caters to dynamic pricing theory in this case, for optimizing revenue, competition analysis and making improvements accordingly.

The pricing of Airbnb listings and sentiment ratings obtained from customer reviews have been found to be significantly correlated by researchers (Wang & Nicolau, 2017). Due to

increased demand and perceived value, positive sentiment ratings frequently permit hosts to charge higher fees, but negative sentiment scores may force hosts to cut costs in an effort to draw more visitors (Zervas, Proserpio, & Byers, 2015). According to certain research, sentiment analysis may also be a factor in dynamic pricing models, assisting hosts in strategically planning price adjustments or enhancements (Park et al., 2019). It is worth noting in order to incorporate more complex features in the models, heavy-duty systems and RAM usage are required, which could be a possible limitation to the study going beyond the realms of explorations into the price prediction and determinants.

Digital profiles, which include elements like online participation, social proof, and digital reputation, are becoming important price factors (Smith, 2021). Hosts with significant online profiles on websites like Instagram or Twitter may use these profiles to support increased charges, especially as social media integration spreads (Johnson & Neuhofer, 2019). Future studies should thus focus on investigating how conventional criteria like location and facilities interact with digital profiles to affect pricing tactics. This will make it possible to grasp Airbnb prices in the digital age's complex mix of factors in a more sophisticated manner (Cheng, 2022).

The Influence of Online Reputation:

One aspect of digital profiles worth exploring further is the influence of online reputation on Airbnb pricing. Hosts who maintain active and positive online reputations through platforms like Instagram and Twitter often have an advantage in attracting guests. Positive online reviews and engagement can create a sense of trust and desirability around a host's listings. As a result, hosts with strong online reputations may have the opportunity to charge higher prices for their properties. However, the relationship between online reputation and pricing is not one-dimensional. Negative online reviews or controversies can also impact a host's ability to command high prices. Hosts must carefully manage their online presence to maintain a positive reputation and, by extension, their ability to set competitive prices.

Exploring the Interplay of Factors:

To gain a comprehensive understanding of Airbnb pricing, future studies should delve deeper into the interplay of various factors. While we have discussed individual predictors such as property type, location, and online reputation, the reality is that these factors often intersect and influence each other. For example, a luxurious property in a prime location with a strong online presence may command exceptionally high prices.

Researchers can employ advanced statistical techniques and machine learning models to unravel the complex relationships among these factors. By doing so, they can provide hosts and investors with more precise insights into how different aspects of their listings and online presence combine to impact pricing decisions. In the realm of Airbnb pricing, the amalgamation of time series analysis and machine learning models has unveiled a panorama of possibilities. The diverse array of predictors, from property attributes to location and digital profiles, collectively shape the pricing landscape. While certain models exhibited remarkable prowess in prediction, the XGBRegressor stood out as the optimal choice, showcasing its adaptability to intricate patterns. Understanding these determinants and their interaction provides not only a roadmap for hosts and investors but also a glimpse into the evolving dynamics of the digital age's pricing ecosystem. As we conclude this section, we recognize that the quest for Airbnb pricing insights continues. The dynamic nature of the sharing economy, coupled with evolving guest preferences and technology, ensures that pricing strategies will remain a vital area of research. The ability to optimize revenue while providing value to guests is a challenge that will persist, and future studies hold the promise of further revelations in the ever-evolving landscape of short-term rentals. The research in this study is subject to several limitations, primarily stemming from the use of aggregated data and a lack of neighborhood-level analysis. These limitations include the potential homogenization of neighborhoods, ignoring spatial autocorrelation, a restricted feature set, and limited consideration of temporal dynamics. This research includes the sentimental analysis and exploration of positive and negative reviews related to neighborhoods and the attributes that justify the price setting is optimal for both guest and hosts.

Future research directions should focus on conducting neighborhood-specific analyses, employing geospatial techniques to visualize spatial distributions, and exploring neighborhood characteristics that influence Airbnb dynamics. Advanced machine learning models and temporal modeling approaches can enhance the accuracy of predictions, while community-based research and an exploration of policy implications can provide a more comprehensive understanding of Airbnb's impact at both city and neighborhood levels. One significant limitation in applying Lancaster's Multi-attribute Theory to Airbnb is the computational power required for image processing and URL mining. While attributes like location, price, and amenities can be easily quantified and analyzed, the processing of visual and web-based elements—such as photographs of the property or external links to local attractions—demands robust computational capabilities. Algorithms for image recognition or natural language processing to parse URLs might require sophisticated machine learning models that are computationally intensive. These technological demands can pose a hurdle, especially for individual hosts or smaller businesses who may not have access to such advanced computational resources, thereby potentially limiting the full realization of attribute-based decision-making and marketing strategies (Chattopadhyay and Mitra, 2019). Airbnb can capitalize on Boston's concerts and events to boost bookings during off-peak seasons. According to Lancaster's Multi-attribute Theory, hosts can offer specialized amenities like event tickets to attract a wider range of tourists (Lancaster, 1966). High-profile events can increase demand, thereby optimizing occupancy rates (Nash, 2021). Such strategies can make Boston a year-round tourist destination, appealing to both domestic and international visitors.
The research was intended to set out Airbnb from the traditional accommodation companies, beyond amenities, bedrooms, bathrooms, room type, property type and include variables like online reviews, cleanliness, accessibility, transit, safety, availability, host details etc. However, more time series based data should be used for various purposes. In order to optimise both occupancy rates and income streams, dynamic pricing may be used in agriculture to modify the cost of farm stays or agritourism experiences based on seasonality, crop cycles, or even weather forecasts. In order to correspond with peak business periods or unique events in a location, Airbnb may implement real-time pricing modifications for office spaces, making it more appealing to

the business clientele(Mahmuda et al., 2022). In the context of concerts, Airbnb may collaborate with event planners to offer dynamic pricing on lodging that is placed strategically close to music venues. Prices can change depending on the popularity of the performance, the availability of tickets, or the star power of the artists. In addition to maximising profits, such a multi-sectoral use of dynamic pricing also improves resource utilisation, generates chances for targeted marketing, and provides value-added services that may address various customer demands and preferences.

RECOMMENDATIONS

Airbnb has also recently announced their plans to enter the commercial real estate market with their new initiative, "Airbnb for Workspaces," which will allow users to book workspace in cities around the world. Given Airbnb's success and continued growth, it's possible that they could explore additional areas of expansion, such as commercial and agricultural property. However, it remains to be seen if and how they will venture into these fields, and whether they will face any challenges or pushback along the way.

CONCLUSION

In the realm of Airbnb pricing, the amalgamation of time series analysis and machine learning models has unveiled a panorama of possibilities. The diverse array of predictors, from property attributes to location and digital profiles, collectively shape the pricing landscape. While certain models exhibited remarkable prowess in prediction, the XGBRegressor stood out as the optimal choice, showcasing its adaptability to intricate patterns. Understanding these determinants and their interaction provides not only a roadmap for hosts and investors but also a glimpse into the evolving dynamics of the digital age's pricing ecosystem. As we conclude this chapter, the quest for Airbnb pricing insights continues, promising further revelations in the ever-evolving landscape of short-term rentals.

REFERENCES

Álvarez-Herranz, A. and Macedo-Ruíz, E., 2021. An evaluation of the three pillars of sustainability in cities with high Airbnb presence: A case study of the city of Madrid. Sustainability, 13(6), p.3220.

Bao, H.X. and Shah, S., 2020. The impact of home sharing on residential real estate markets. Journal of Risk and Financial Management, 13(8), p.161.

Benckendorff, P.J., Xiang, Z. and Sheldon, P.J., 2019. Tourism information technology. Cabi.

Breiman, L., 2001. Random forests. Machine learning, 45, pp.5-32.

Cansoy, M.S., 2018. " Sharing" in Unequal Spaces: Short-term Rentals and the Reproduction of Urban Inequalities (Doctoral dissertation, Boston College).

Chapman, S., Mohammad, S. and Villegas, K., 2023. Predicting Listing Prices In Dynamic Short Term Rental Markets Using Machine Learning Models. arXiv preprint arXiv:2308.06929.

Chattopadhyay, M. and Mitra, S.K., 2019. Do airbnb host listing attributes influence room pricing homogenously?. International Journal of Hospitality Management, 81, pp.54-64.

Chiny, M., Bencharef, O., Hadi, M.Y. and Chihab, Y., 2021. A client-centric evaluation system to evaluate guest's satisfaction on AirBNB using machine learning and NLP. Applied Computational Intelligence and Soft Computing, 2021, pp.1-14.

Chornous, G. and Horbunova, Y., 2020. Modeling and Forecasting Dynamic Factors of Pricing in E-commerce. In IT&I (pp. 71-82).

Coles, P.A., Egesdal, M., Ellen, I.G., Li, X. and Sundararajan, A., 2017. Airbnb usage across New York City neighborhoods: Geographic patterns and regulatory implications. Forthcoming, Cambridge Handbook on the Law of the Sharing Economy.

Cover, T. and Hart, P., 1967. Nearest neighbor pattern classification. IEEE transactions on information theory, 13(1), pp.21-27.

Davidson, N.M. and Infranca, J.J., 2015. The sharing economy as an urban phenomenon. Yale L. & Pol'y Rev., 34, p.215.

De Jaureguizar Cervera, D., Pérez-Bustamante Yábar, D.C. and de Esteban Curiel, J., 2022. Factors affecting short-term rental first price: A revenue management model. Frontiers in Psychology, 13, p.994910.

Domènech, A., Larpin, B., Schegg, R. and Scaglione, M., 2019. Disentangling the geographical logic of Airbnb in Switzerland. Erdkunde, (H. 4), pp.245-258.

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Annals of statistics, pp.1189-1232.

Freund, Y. and Schapire, R.E., 1997. A decision-theoretic generalization of on-line learning and an application to boosting. Journal of computer and system sciences, 55(1), pp.119-139.

Garcia, S.G., 2023. Evaluating the Impact of Image Features on Airbnb Price Predictions: A Machine Learning Approach to Hedonic Pricing (Master's thesis).

Gibbs, C., Guttentag, D., Gretzel, U., Morton, J. and Goodwill, A., 2018. Pricing in the sharing economy: A hedonic pricing model applied to Airbnb listings. Journal of Travel & Tourism Marketing, 35(1), pp.46-56.

Glaeser, E.L., Kominers, S.D., Luca, M. and Naik, N., 2018. Big data and big cities: The promises and limitations of improved measures of urban life. Economic Inquiry, 56(1), pp.114-137.

Guttentag, D., 2015. Airbnb: disruptive innovation and the rise of an informal tourism accommodation sector. Current issues in Tourism, 18(12), pp.1192-1217.

Guttentag, D., 2016. Why tourists choose Airbnb: A motivation-based segmentation study underpinned by innovation concepts.

Guttentag, D., Smith, S., Potwarka, L. and Havitz, M., 2018. Why tourists choose Airbnb: A motivation-based segmentation study. Journal of Travel Research, 57(3), pp.342-359.

Ho, C.I., Chen, T.S. and Li, C.P., 2023. Airbnb's Negative Externalities from the Consumer's Perspective: How the Effects Influence the Booking Intention of Potential Guests. Sustainability, 15(11), p.8695.

Jiang, Y., Zhang, H., Cao, X., Wei, G. and Yang, Y., 2023. How to better incorporate geographic variation in Airbnb price modeling? Tourism Economics, 29(5), pp.1181-1203.

Kwok, L. and Xie, K.L., 2019. Pricing strategies on Airbnb: Are multi-unit hosts revenue pros?. International Journal of Hospitality Management, 82, pp.252-259.

Liang, L.J., 2015. Understanding repurchase intention of Airbnb consumers: perceived authenticity, EWoM and price sensitivity (Doctoral dissertation, University of Guelph).

Liang, L.J., Choi, H.C. and Joppe, M., 2018. Understanding repurchase intention of Airbnb consumers: perceived authenticity, electronic word-of-mouth, and price sensitivity. Journal of Travel & Tourism Marketing, 35(1), pp.73-89.

Lima, V., 2019. Towards an understanding of the regional impact of Airbnb in Ireland. Regional Studies, Regional Science, 6(1), pp.78-91.

Lladós-Masllorens, J., Meseguer-Artola, A. and Rodríguez-Ardura, I., 2020. Understanding peer-to-peer, two-sided digital marketplaces: pricing lessons from Airbnb in Barcelona. Sustainability, 12(13), p.5229.

Ma, X., Hancock, J.T., Lim Mingjie, K. and Naaman, M., 2017, February. Self-disclosure and perceived trustworthiness of Airbnb host profiles. In Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing (pp. 2397-2409).

Mahmuda, S., Sigler, T., Corcoran, J. and Knight, E., 2022. Airbnb and micro-entrepreneurship in regional economies: Lessons from Australia. Geographical Research, 60(2), pp.269-285.

Mora-Garcia, R.T., Cespedes-Lopez, M.F. and Perez-Sanchez, V.R., 2022. Housing Price Prediction Using Machine Learning Algorithms in COVID-19 Times. Land, 11(11), p.2100.

Nash, K., 2021. Data Visualization Can Shifts our Sharing Economy Perceptions: Austin, Texas Airbnb Landscape. In *AMCIS*.

Nunes, M.R.D.S.P., 2023. Predicting and explaining Airbnb prices in Lisbon: machine learning approach (Doctoral dissertation).

Oskam, J. and Boswijk, A., 2016. Airbnb: the future of networked hospitality businesses. Journal of tourism futures, 2(1), pp.22-42.

Oskam, J., van der Rest, J.P. and Telkamp, B., 2018. What's mine is yours—but at what price? Dynamic pricing behavior as an indicator of Airbnb host professionalization. Journal of Revenue and Pricing Management, 17, pp.311-328.

Pappas, N., 2017. The complexity of purchasing intentions in peer-to-peer accommodation. International Journal of Contemporary Hospitality Management, 29(9), pp.2302-2321.

Palomino, M., Taylor, T. and Gössling, S., 2016. The sharing economy: Comparing ride-hailing and accommodation platforms in Latin America. Current issues in Tourism, 19(12), pp.1301-1308.

Park, S., Wang, X., Xia, J. and Choi, H.S., 2019. Airbnb adoption by travelers: Refinement of the technology acceptance model. Journal of Travel Research, 0047287519889282.

Patil, N., Bouchout, R., Douma, F. and Hawkins, P., 2018. Regulatory responses to the sharing economy: an analysis of Airbnb in American cities. Land Use Policy, 73, pp.284-291.

Roberts, B., Tam, G., Choi, J., Stone, J., Vaugier, L. and DiRico, M., 2019. A random forest machine learning approach for predicting price of short-term vacation rental properties. Journal of Retailing and Consumer Services, 50, pp.212-219.

Su, Q., Shi, B., Zheng, Y. and Wang, W., 2019. Forecasting the short-term room price for Airbnb listings in Shanghai. Tourism Economics, 25(2), pp.195-214.

Suthaharan, S., 2014. Big data classification: Problems and challenges in network intrusion prediction with machine learning. *ACM SIGMETRICS Performance Evaluation Review*, *41*(4), pp.70-73.

Wang, D., Li, X. and Li, Y., 2013. China's "smart tourism destination" initiative: A taste of the service-dominant logic. Journal of Destination Marketing & Management, 2(2), pp.77-81.

Wang, X., Chi, C.G.Q. and Li, X., 2019. China's "smart tourism destination" initiative: A taste of the service-dominant logic. Journal of Destination Marketing & Management, 12, pp.77-81.

Wielechowski, M., Zavatskaya, A. and Pukala, R., 2021. The Impact of COVID-19 on the Airbnb Market: The Case of Locations in Poland. Sustainability, 13(11), p.5812.

Wirtz, J. and Tang, C., 2016. Airbnb: its working model and the monetary aspects of sharing. Journal of Revenue and Pricing Management, 15(6), pp.509-515.

Xie, K., 2019. Impacts of the sharing economy on the dynamics of hotel room pricing: A case study of Airbnb and hotels in San Francisco. Journal of Hospitality and Tourism Management, 38, pp.89-98.

Zervas, G., Proserpio, D. and Byers, J.W., 2017. The rise of the sharing economy: Estimating the impact of Airbnb on the hotel industry. Journal of Marketing Research, 54(5), pp.687-705.

Zhang, S., Yao, L., Sun, A. and Tay, Y., 2019. Deep learning based recommender system: A survey and new perspectives. *ACM computing surveys (CSUR)*, *52*(1), pp.1-38.

Appendix

1	host_id	3.324573802971576
2	host_listings_count	Infinity
3	host_total_listings_count	Infinity
4	neighbourhood_group_cleansed	NaN
5	latitude	1571066.4706374314
6	longitude	1570787.4483485115
7	accommodates	15.425486343152592
8	bathrooms	8.648131036306516
9	bedrooms	9.937226223654036
10	beds	12.2004750406324
11	square_feet	1.0844155526471944
12	guests_included	4.100109755730556
13	minimum_nights	1.161412731698916
14	maximum_nights	1.0073420269274724
15	has_availability	NaN
16	availability_30	19.75498985019606
17	availability_60	79.13474899641307
18	availability_90	51.12172069527602
19	availability_365	5.139829370855178
20	number_of_reviews	2.888655561916803
21	review_scores_rating	205.09554425195736
22	review_scores_accuracy	169.44761269217875
23	review_scores_cleanliness	142.02186100746255

Activate Windows
Go to Settings to activate Windows.

+ Code + Text All changes saved

4	neighbourhood_group_cleansed	NaN
5	latitude	1571066.4706374314
6	longitude	1570787.4483485115
7	accommodates	15.425486343152592
8	bathrooms	8.648131036306516
9	bedrooms	9.937226223654036
10	beds	12.2004750406324
11	square_feet	1.0844155526471944
12	guests_included	4.100109755730556
13	minimum_nights	1.161412731698916
14	maximum_nights	1.0073420269274724
15	has_availability	NaN
16	availability_30	19.75498985019606
17	availability_60	79.13474899641307
18	availability_90	51.12172069527602
19	availability_365	5.139829370855178
20	number_of_reviews	2.888655561916803
21	review_scores_rating	205.09554425195736
22	review_scores_accuracy	169.44761269217875
23	review_scores_cleanliness	142.02186100746255
24	review_scores_checkin	238.88201505272667

Show 25 per page

Activate Windows

```
# Display the VIF scores as a visual table
vif_scores
```

```
/usr/local/lib/python3.10/dist-packages/statsmodels/stats/outliers_influence.py:198: RuntimeWarning:

divide by zero encountered in double_scalars

/usr/local/lib/python3.10/dist-packages/statsmodels/regression/linear_model.py:1783: RuntimeWarning:

invalid value encountered in double_scalars
```

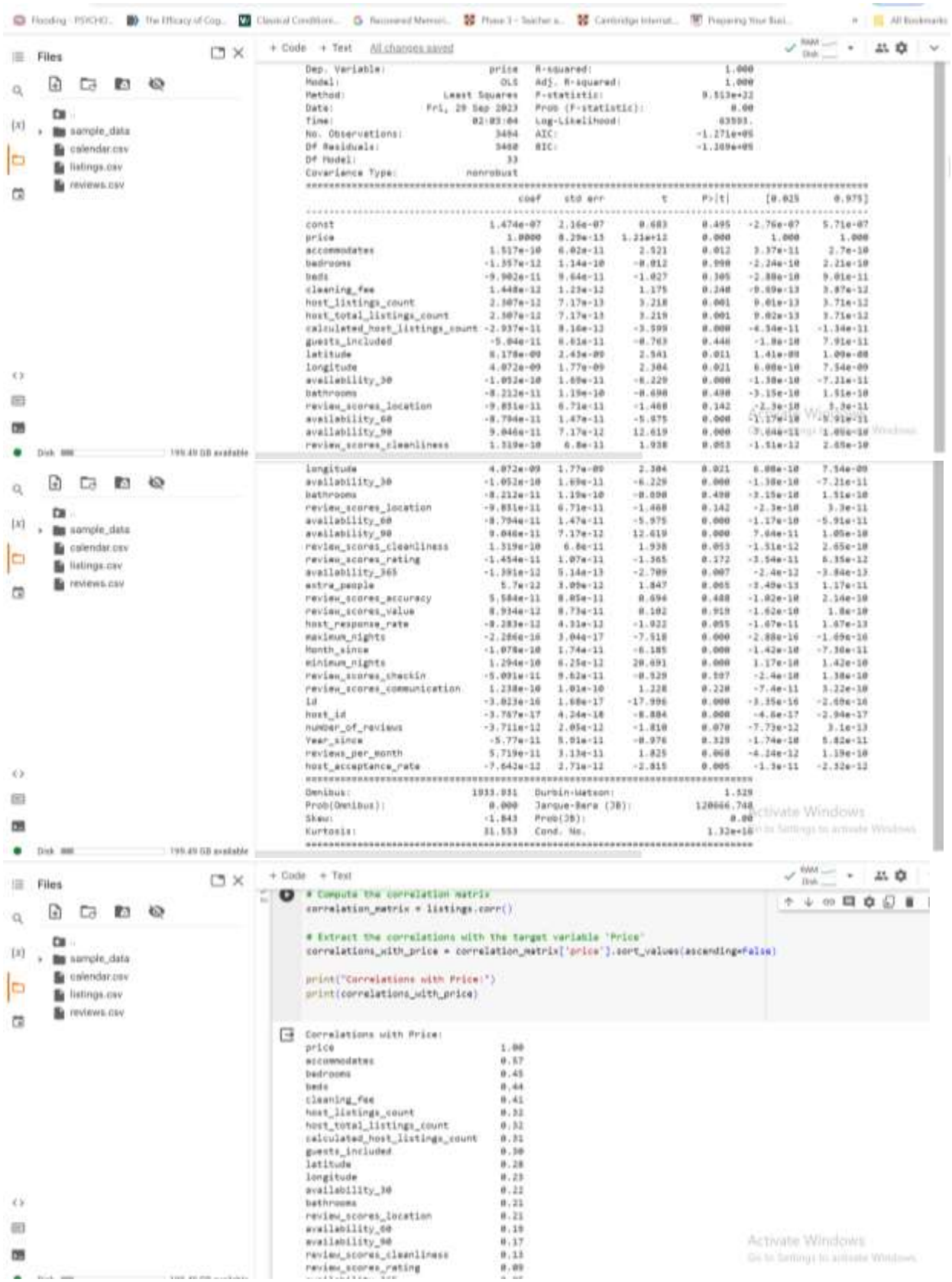
26 to 30 of 30 entries Filter

index	Variable	VIF
25	review_scores_communication	244.28035867136785
26	review_scores_location	110.39909180149638
27	review_scores_value	193.5466766224511
28	calculated_host_listings_count	22.228326695518074
29	reviews_per_month	3.4736167493166348

Show 25 per page

Like what you see? Visit the data table notebook to learn more about interactive tables.

Activate Windows



Files

+ Code + Text All changes saved

sample_data
calendar.csv
listings.csv
reviews.csv

```
Dep. Variable:                  price   R-squared:                       1.000
Model:                            OLS   Adj. R-squared:                  1.000
Method:                 Least Squares   F-statistic:                 9.513e+32
Date:                Fri, 29 Sep 2023   Prob (F-statistic):               0.00
Time:                        02:03:04   Log-Likelihood:                 63593.
No. Observations:                3494   AIC:                        -1.271e+05
Df Residuals:                    3460   BIC:                        -1.269e+05
Df Model:                          33
Covariance Type:            nonrobust
==================================================================================================
                                     coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------------
const                           1.474e-07   2.16e-07      0.683      0.495   -2.76e-07    5.71e-07
price                              1.0000   8.29e-13   1.21e+12      0.000       1.000       1.000
accommodates                    1.517e-10   6.02e-11      2.521      0.012    3.37e-11    2.7e-10
bedrooms                       -1.357e-12   1.14e-10     -0.012      0.990   -2.24e-10    2.21e-10
beds                           -9.902e-11   9.64e-11     -1.027      0.305   -2.88e-10    9.01e-11
cleaning_fee                    1.448e-12   1.23e-12      1.175      0.240   -9.69e-13    3.87e-12
host_listings_count             2.307e-12   7.17e-13      3.218      0.001    9.01e-13    3.71e-12
host_total_listings_count       2.307e-12   7.17e-13      3.219      0.001    9.02e-13    3.71e-12
calculated_host_listings_count -2.937e-11   8.16e-12     -3.599      0.000   -4.54e-11   -1.34e-11
guests_included                -5.04e-11   6.61e-11     -0.763      0.446    -1.8e-10    7.91e-11
latitude                        6.178e-09   2.43e-09      2.541      0.011    1.41e-09    1.09e-08
longitude                       4.872e-09   1.77e-09      2.304      0.021    6.88e-10    7.54e-09
availability_30                -1.052e-10   1.69e-11     -6.229      0.000   -1.38e-10   -7.21e-11
bathrooms                      -8.212e-11   1.19e-10     -0.690      0.490   -3.15e-10    1.51e-10
review_scores_location         -9.851e-11   6.71e-11     -1.468      0.142    -2.3e-10    3.3e-11
availability_60                -8.794e-11   1.47e-11     -5.975      0.000   -1.17e-10   -5.91e-11
availability_90                 9.046e-11   7.17e-12     12.619      0.000    7.64e-11    1.05e-10
review_scores_cleanliness       1.319e-10    6.8e-11      1.938      0.053   -1.51e-12    2.65e-10
```

```
longitude                       4.872e-09   1.77e-09      2.304      0.021    6.88e-10    7.54e-09
availability_30                -1.052e-10   1.69e-11     -6.229      0.000   -1.38e-10   -7.21e-11
bathrooms                      -8.212e-11   1.19e-10     -0.690      0.490   -3.15e-10    1.51e-10
review_scores_location         -9.851e-11   6.71e-11     -1.468      0.142    -2.3e-10    3.3e-11
availability_60                -8.794e-11   1.47e-11     -5.975      0.000   -1.17e-10   -5.91e-11
availability_90                 9.046e-11   7.17e-12     12.619      0.000    7.64e-11    1.05e-10
review_scores_cleanliness       1.319e-10    6.8e-11      1.938      0.053   -1.51e-12    2.65e-10
review_scores_rating           -1.454e-11   1.07e-11     -1.365      0.172   -3.54e-11    6.35e-12
availability_365               -1.391e-12   5.14e-13     -2.709      0.007    -2.4e-12   -3.84e-13
extra_people                    5.7e-12    3.09e-12      1.847      0.065   -3.49e-13    1.17e-11
review_scores_accuracy          5.584e-11   8.05e-11      0.694      0.488   -1.02e-10    2.14e-10
review_scores_value             8.934e-12   8.73e-11      0.102      0.919   -1.62e-10    1.8e-10
host_response_rate             -8.283e-12   4.31e-12     -1.922      0.055   -1.67e-11    1.67e-13
maximum_nights                 -2.286e-16   3.04e-17     -7.518      0.000   -2.88e-16   -1.69e-16
Month_since                    -1.078e-10   1.74e-11     -6.185      0.000   -1.42e-10   -7.36e-11
minimum_nights                  1.294e-10   6.25e-12     20.691      0.000    1.17e-10    1.42e-10
review_scores_checkin          -5.091e-11   9.62e-11     -0.529      0.597    -2.4e-10    1.38e-10
review_scores_communication     1.238e-10   1.01e-10      1.228      0.220    -7.4e-11    3.22e-10
id                             -3.823e-16   1.68e-17    -17.996      0.000   -3.35e-16   -2.69e-16
host_id                        -3.767e-17   4.24e-18     -8.884      0.000    -4.6e-17   -2.94e-17
number_of_reviews              -3.711e-12   2.05e-12     -1.810      0.070   -7.73e-12    3.1e-13
Year_since                      -5.77e-11   5.01e-11     -0.976      0.329    -1.74e-10    5.82e-11
reviews_per_month               5.719e-11   3.13e-11      1.825      0.068    -4.14e-12    1.19e-10
host_acceptance_rate           -7.643e-12   2.71e-12     -2.815      0.005    -1.3e-11   -2.32e-12
==================================================================================================
Omnibus:                     1933.931   Durbin-Watson:                   1.529
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           120666.748
Skew:                          -1.843   Prob(JB):                         0.00
Kurtosis:                      31.553   Cond. No.                     1.32e+16
==================================================================================================
```

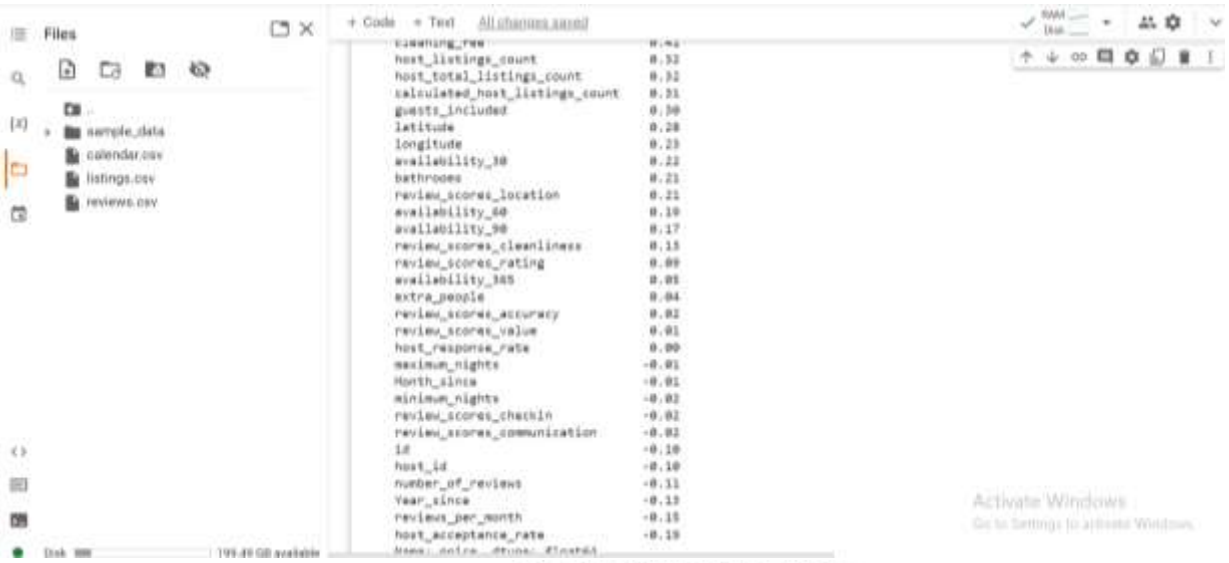
Files

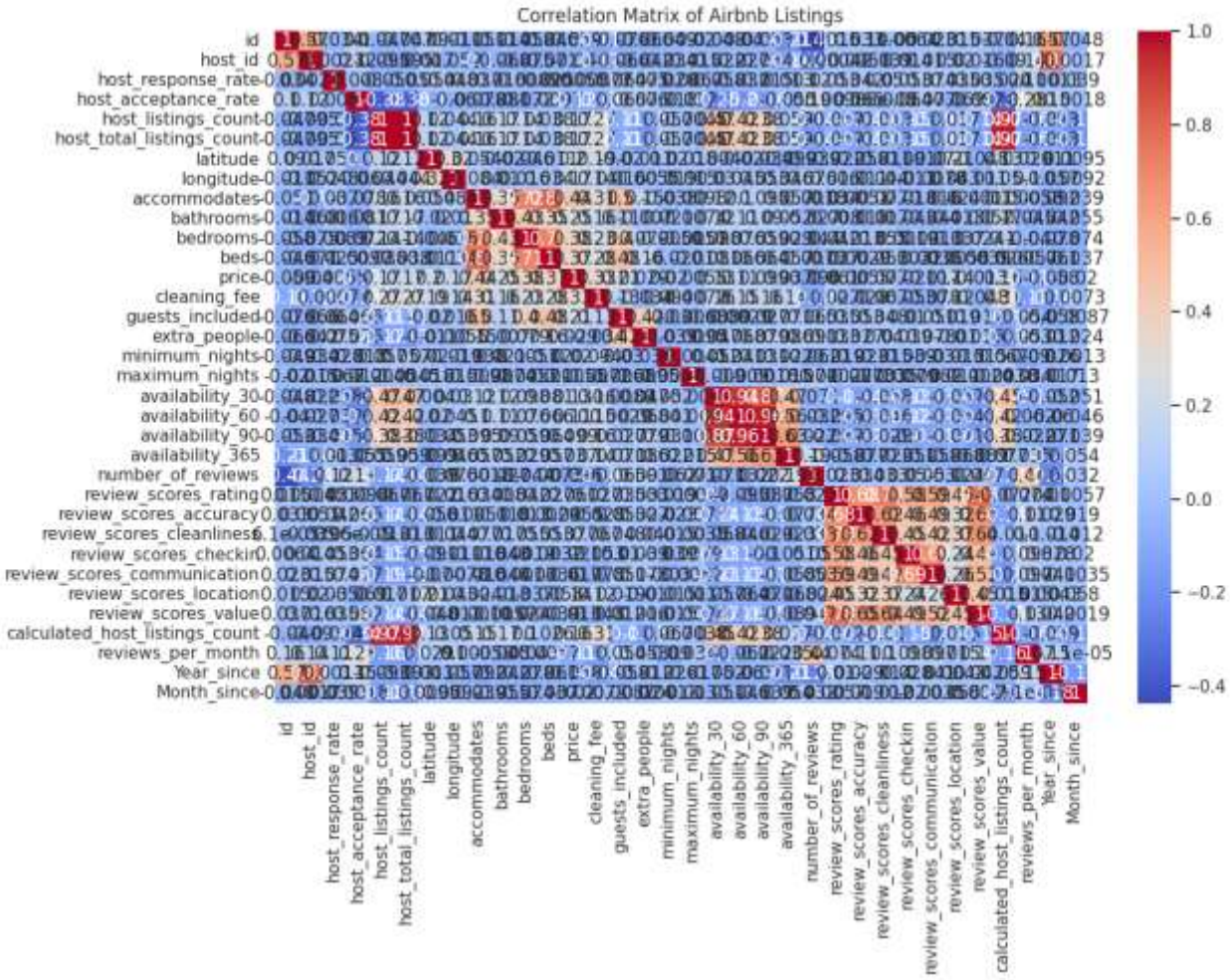
+ Code + Text

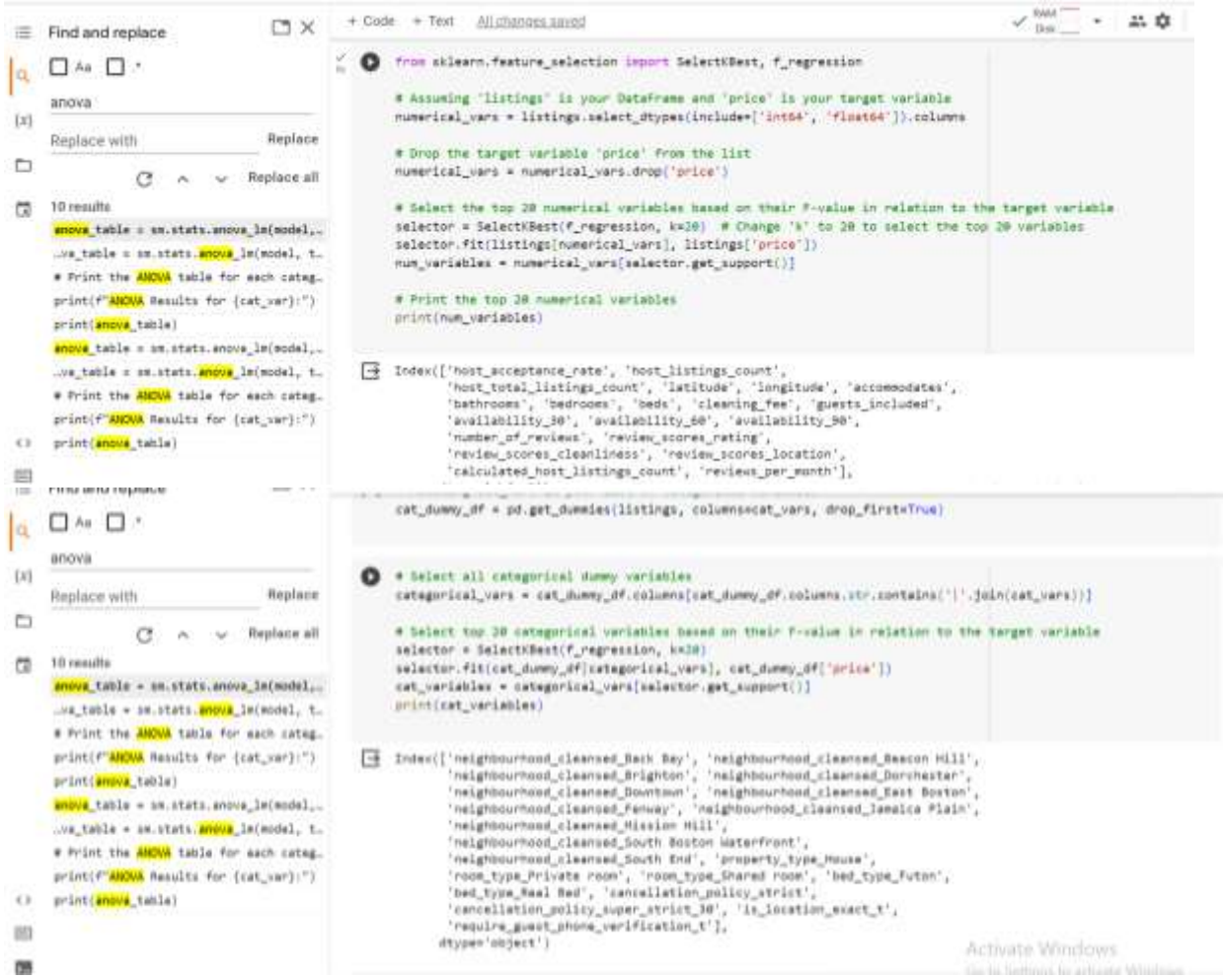
sample_data
calendar.csv
listings.csv
reviews.csv

```
# Compute the correlation matrix
correlation_matrix = listings.corr()

# Extract the correlations with the target variable 'Price'
correlations_with_price = correlation_matrix['price'].sort_values(ascending=False)

print("Correlations with Price:")
print(correlations_with_price)
```

```
Correlations with Price:
price                           1.00
accommodates                    0.57
bedrooms                        0.45
beds                            0.44
cleaning_fee                    0.41
host_listings_count             0.31
host_total_listings_count       0.31
calculated_host_listings_count  0.31
guests_included                 0.30
latitude                        0.28
longitude                       0.23
availability_30                 0.22
bathrooms                       0.21
review_scores_location          0.21
availability_60                 0.19
availability_90                 0.17
review_scores_cleanliness       0.13
review_scores_rating            0.09
```


Files
sample_data
calendar.csv
listings.csv
reviews.csv
+ Code
+ Text
All changes saved
host_listings_count 0.53
host_total_listings_count 0.32
calculated_host_listings_count 0.31
guests_included 0.30
latitude 0.28
longitude 0.23
availability_30 0.22
bathrooms 0.21
review_scores_location 0.21
availability_60 0.19
availability_90 0.17
review_scores_cleanliness 0.15
review_scores_rating 0.09
availability_365 0.05
extra_people 0.04
review_scores_accuracy 0.02
review_scores_value 0.01
host_response_rate 0.00
maximum_nights -0.01
Month_since -0.01
minimum_nights -0.02
review_scores_checkin -0.02
review_scores_communication -0.02
id -0.10
host_id -0.10
number_of_reviews -0.11
Year_since -0.13
reviews_per_month -0.15
host_acceptance_rate -0.19
Activate Windows


Correlation Matrix of Airbnb Listings
id
host_id
host_response_rate
host_acceptance_rate
host_listings_count
host_total_listings_count
latitude
longitude
accommodates
bathrooms
bedrooms
beds
price
cleaning_fee
guests_included
extra_people
minimum_nights
maximum_nights
availability_30
availability_60
availability_90
availability_365
number_of_reviews
review_scores_rating
review_scores_accuracy
review_scores_cleanliness
review_scores_checkin
review_scores_communication
review_scores_location
review_scores_value
calculated_host_listings_count
reviews_per_month
Year_since
Month_since
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4



Declaration

i. I affirm the content of this thesis is solely the result of my own efforts;

ii. All external sources and materials have been appropriately credited within the document;

iii. None of the work presented in this thesis has been submitted to secure any other degree or certification from this or any other educational institution.