Airbnb in Denmark: A Deep Dive into Sealand's Properties

Anh Vi Pham (Henry)

20-11-2023

Table of Contents

Executive summary

1. **Problem statement**

   Utilizing Airbnb property data in Denmark, this project aims to delve into a dataset comprising 23,941 rentals spanning from November 2016 to October 2019. The objective is to extract insights and present a comprehensive overview of the housing market. Given that Sealand encompasses the capital city, Copenhagen, the analysis will specifically explore factors distinguishing properties in Sealand from the rest of Denmark. Additionally, the project includes the development of a predictive model to classify properties based on their price.

2. **Insights**

   Based on the analysis, several observations have been made:

   - The nightly Airbnb rates in Sealand are slightly higher than the rest of Denmark, but the prices vary significantly. On the other hand, there are many outliers in the prices of Airbnb properties outside Sealand.
   - Over the three-year period from November 2016, the average prices in both areas show an increasing trend. Particularly in Sealand, there is a recurring price cycle, with a price dip usually occurring in January each year before gradually rising again.
   - The location of the property is the most influential factor in price fluctuations, alongside other factors such as room type and the number of occupants allowed in a rental.
   - There is sufficient evidence, at a 95% confidence level, to conclude that the average satisfaction of properties in Sealand is higher than the rest of Denmark.
   - 5 is the recommended number of properties that a landlord should own to ensure the quality of their Airbnb services.

Data source

1. **Data source**
   The data pertains to properties listed on Airbnb in Denmark during the period 2016-2019, provided by Deakin University for educational purposes.

2. **Meta data**
   Explaination for attributes in the data:
   - prop_room_id: Property / room id, its type and price per minimum nights' stay (in US$)
   - host_id: id of the property host (the person owning it or renting it out)
   - prop_room_type: type of room
   - Bedrooms, description, etc.
   - neighborhood: Property geo-location and its neighbourhood
   - reviews: number of reviews recorded so far (since listing)
   - overall_satisfaction: Overall satisfaction (average from all people who have rented the property)
   - accommodates: The number of occupants allowed in a rental
   - price_USD: price of property
   - minstay: Minimum number of nights to be booked (if applicable)
   - description: description of the property by its host.
   - latitude: Latitude of property
   - longtitude: Longitude of property
   - first-listed: Date first listed

Data exploration, pattern discovery, and preparation

1. Data preparing

Data preprocessing techniques were applied for the analysis:

- Missing values: the data contains 4 missing values in two variables: 'accommodates' and 'price_USD' which were dropped. Besides, 86 missing values were also found in 'description' but since the analysis did not involve in text analysis, they were kept intact.
- Duplicated values: no duplicated value was found.
- Specifying properties in Sealand: using 'latitude' and 'longitude', such properties were identified relatively.

2. Data exploration

a. Data distribution

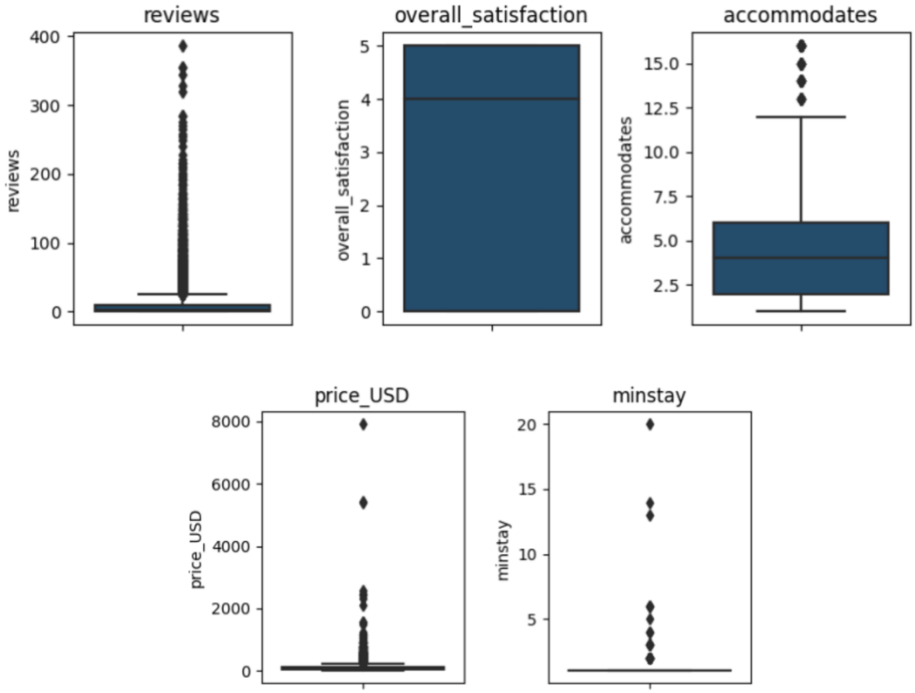

Figure 1: Data distribution of 5 numerical variables.

From the plots, we can draw two main observations:

- Most of the variables exhibit a strong right-skewness.
- Numerous outliers are noticeable, particularly in the case of the number of reviews and prices.

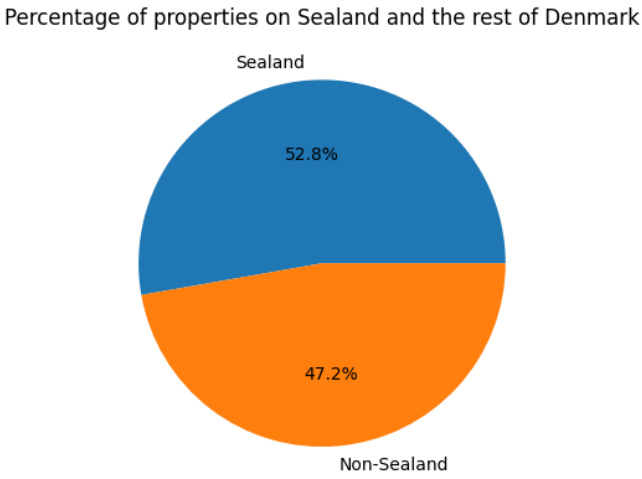

Figure 2: Proportion of properties located in Sealand and the rest of Denmark.

Among the 23,936 properties included in the dataset, 52.8% are situated in Sealand, totaling 12,630 properties.

b. Correlation Matrix

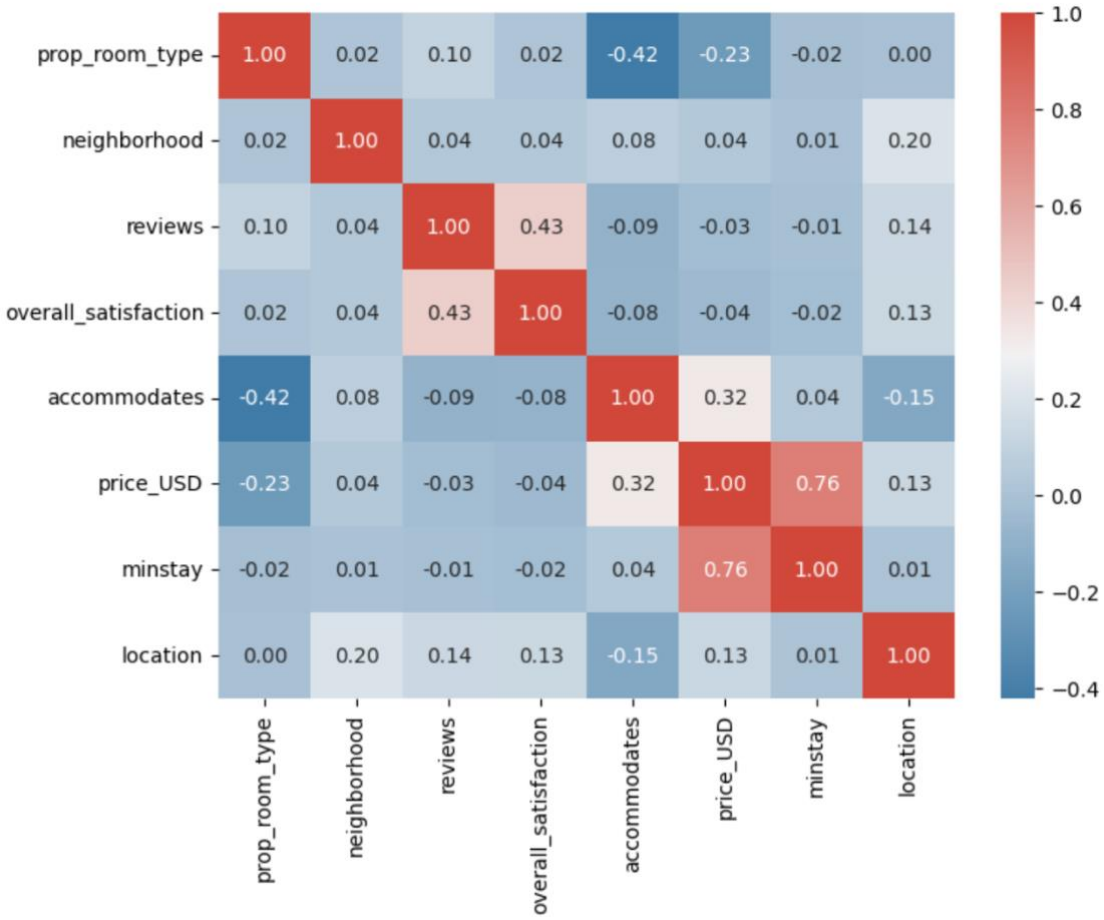

Figure 3: Correlation matrix.

The variable "accommodates" shows a high correlation with both "prop_room_type" and "price_USD." This indicates that larger room types can accommodate more people and tend to have higher associated costs.

3. Sealand's properties analysis

As Sealand incorporates the capital city of Copenhagen (Danish: København), AirBNB wants to know if there are differences between the Sealand properties versus those in the rest of Denmark. Explore this from the perspective of people staying at the rentals (define)

a. Price comparison for properties in Sealand compared to the rest of Denmark

- **Overall price**

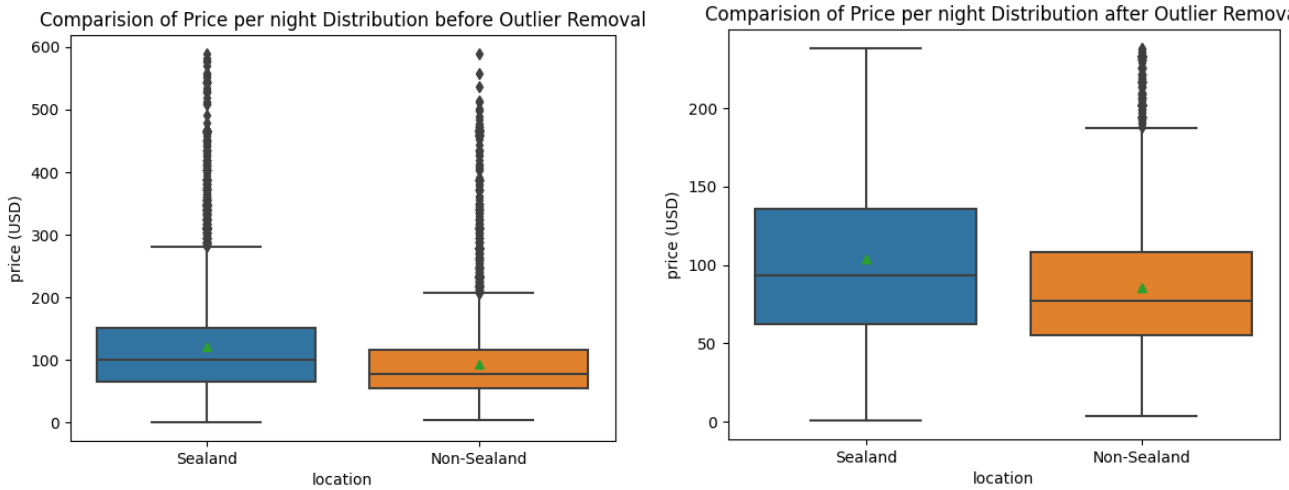

Figure 4: Price comparision of properties in Sealand and the rest of Denmark.

With Figure 4, some observations can be given as follow:

- Based on the median, we can see that the overall price of property on Sealand is slightly higher than the rest of Denmark.
- However, the range (IQR) of the price of property on Sealand is broader, indicating an inconsistency on the contrary with non-Sealand.
- Noticeably, there are multiple outliers for the price per night of properties on the rest of Denmark.

- **Average price over time**

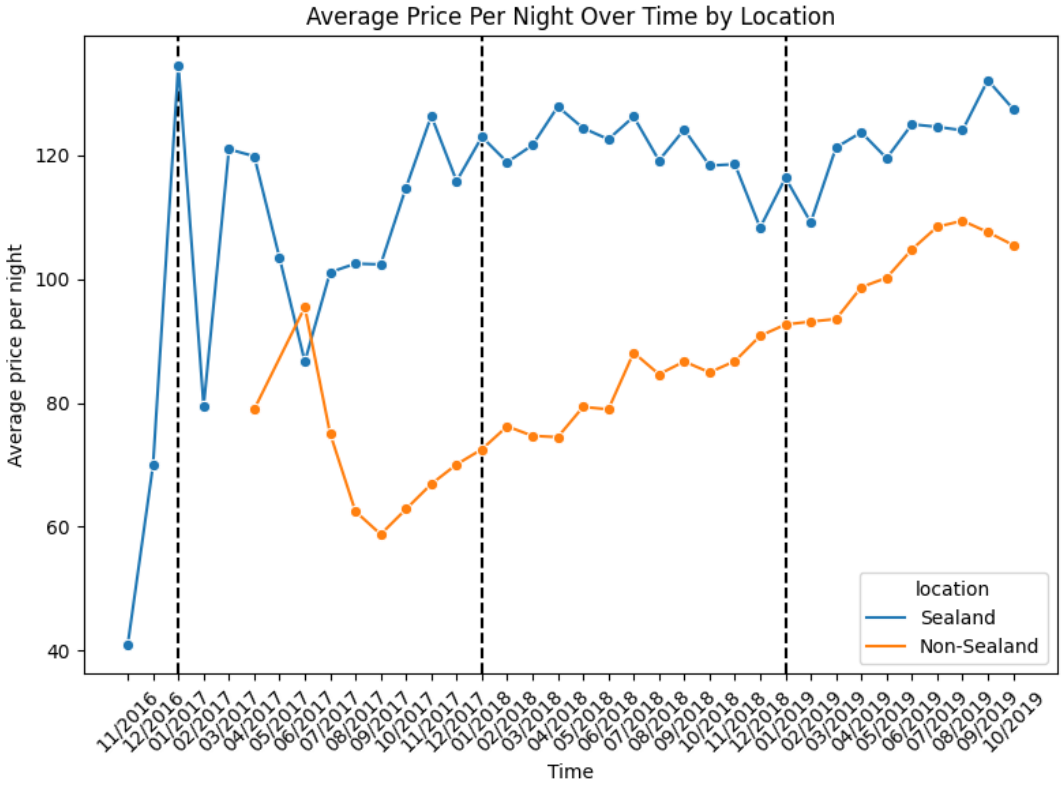

Figure 5: Average price per night overtime by location.

From the plot, we can observe that, on average, the price of properties located in Sealand was higher compared to that of non-Sealand for most of the given period, except for a brief period in May 2017. Furthermore:

- The overall trend of the average property prices in both areas was consistently upward.
- The data indicates a slightly seasonal nature for Sealand's average property prices, occurring annually. Prices experienced a significant drop in January each year before undergoing a considerable rise immediately afterward. The average price then exhibited an unstable upward trend before experiencing a slight drop in October each year.
- Price estimating with Linear regression analysis.

- **Price estimation with Linear Regression Analysis**

OLS Regression Results

Dep. Variable:	price_per_night	R-squared:	0.296
Model:	OLS	Adj. R-squared:	0.296
Method:	Least Squares	F-statistic:	2009.
Date:	Wed, 15 Nov 2023	Prob (F-statistic):	0.00
Time:	11:10:59	Log-Likelihood:	-1.0632e+05
No. Observations:	19148	AIC:	2.126e+05
Df Residuals:	19143	BIC:	2.127e+05
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	42.6452	1.389	30.705	0.000	39.923	45.367
prop_room_type	-27.0400	1.098	-24.626	0.000	-29.192	-24.888
overall_satisfaction	-1.3385	0.192	-6.985	0.000	-1.714	-0.963
accommodates	13.7482	0.223	61.759	0.000	13.312	14.185
location	38.0173	0.923	41.207	0.000	36.209	39.826

Omnibus:	9501.866	Durbin-Watson:	2.005
Prob(Omnibus):	0.000	Jarque-Bera (JB):	89115.821
Skew:	2.187	Prob(JB):	0.00
Kurtosis:	12.621	Cond. No.	19.1

After several iterations of removing attributes with multicollinearity and those that did not have significant statistical influence on the target attribute, the model is now more refined and completed.

Equation: price_per_night = -27.04*prop_room_type - 1.3385*overall_satisfaction + 13.7482*accommodates + 38.0173*location

Among the various attributes in the model, location stands out as the most prominent factor. On average, the price for a one-night stay for properties located in Sealand will be $38 higher compared to properties in the rest of Denmark.

According to the R-squared, on average, 29.6% of the variation in price_per_night can be explained by the variation in the 4 included attributes. The remaining 70.4% of variation would be explained by other factors not mentioned in the model. Hence, the model is not confident enough when it comes to explaining the variation in price_per_night.

b. Customer Satisfaction comparison for properties in Sealand compared to the rest of Denmark

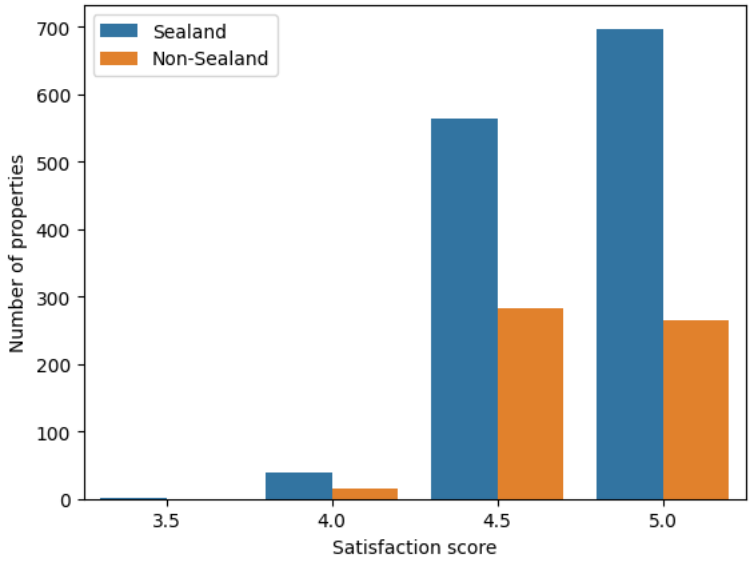

Figure 6: Satisfaction score distribution by Location

Based on the number of properties in each rating category (Figure 6), it can be observed that:

- The data lacks information about properties with low overall_satisfaction (below 3 points).
- Properties with sufficient data mostly achieve very high overall_satisfaction ratings.
- The number of properties with high satisfaction levels in Sealand is significantly higher than the rest of Denmark.

A hypothesis test was conducted to examine whether the satisfaction scores of properties located in Sealand were indeed higher than those in the rest of Denmark. Before delving into the results, let's take a look at the distribution of the data (refer to Figure 6 and 7):

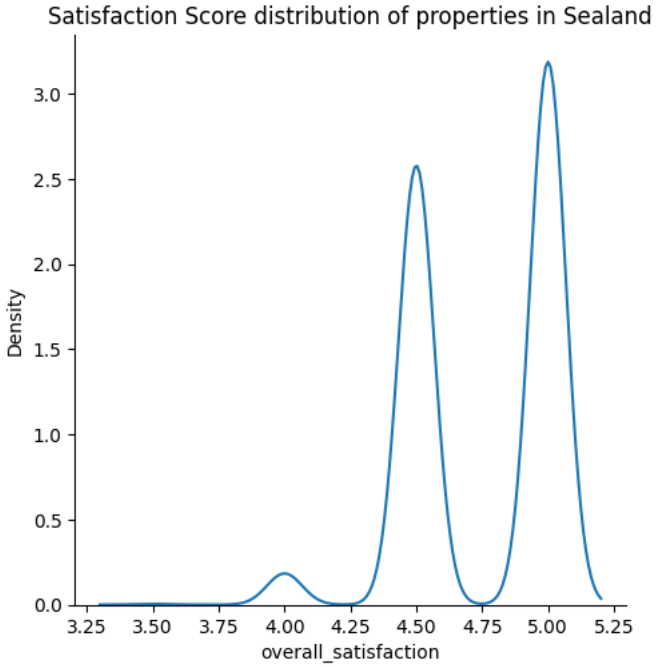

Figure 6: Satisfaction score distribution of properties in Sealand.

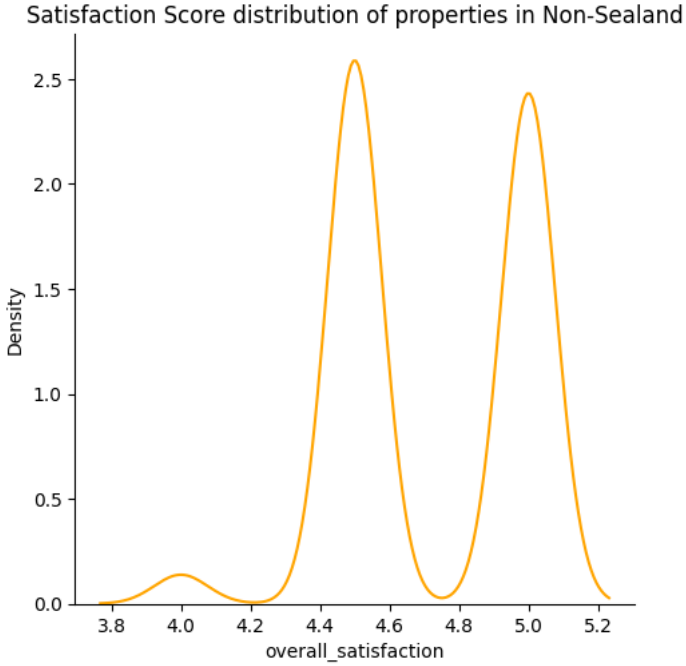

Figure 7: Satisfaction score distribution of properties in the rest of Denmark.

Observing the plots, it is evident that neither dataset follows a normal distribution. Given this, the Mann-Whitney U test was employed as the hypothesis testing method for this task.

Testing process and result:

- Hypothesis statement:
  - H0: average satisfaction score of properties in Sealand <= average satisfaction score of properties in the rest of Denmark
  - H1: average satisfaction score of properties in Sealand > average satisfaction score of properties in the rest of Denmark
- alpha: 0.05
- statistic: 387645.5
- p-value: 0.009921416231619034

Conclusion: There is sufficient evidence, at 95% confidence level, to conclude the average satisfaction of properties in Sealand is higher than the rest of Denmark.

c. Landlords in Sealand

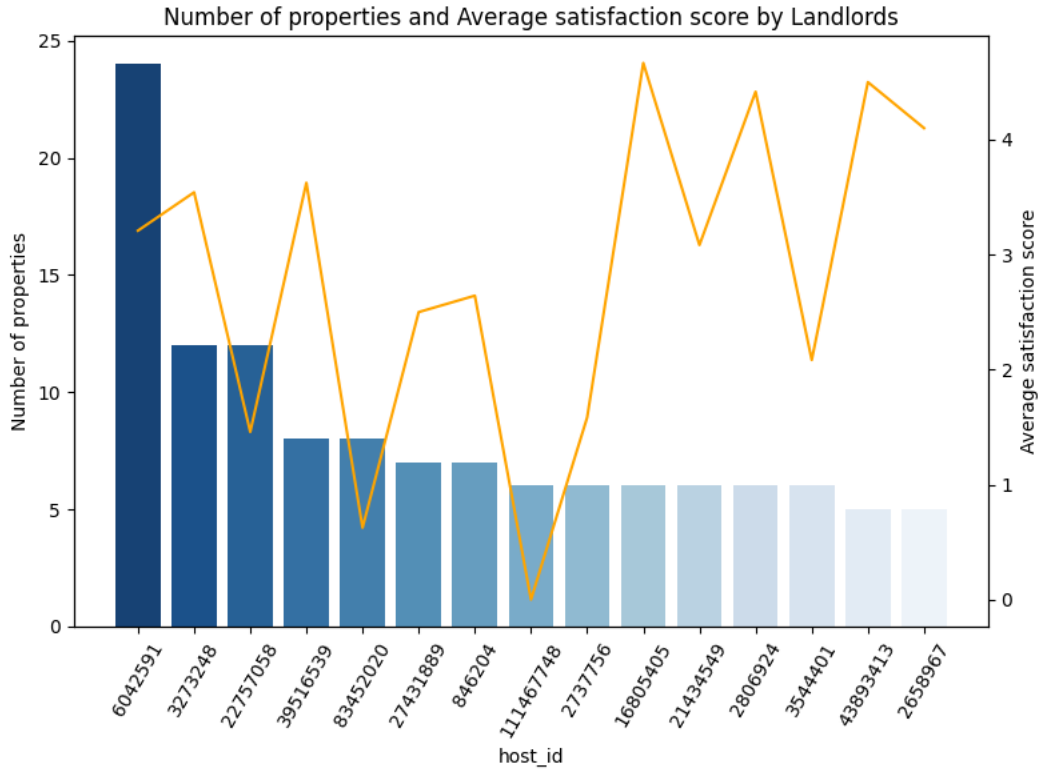

Figure 8: Number of properties and Average satisfaction score by Landlords

From the plot, it is evident that the host with ID 6042591 possesses the highest number of properties, totaling approximately 25 houses. However, hosts with more than 6 properties seem to exhibit a lower average satisfaction score. Based on these insights, it is recommended that landlords consider owning around 5-6 properties to better meet customer expectations.

I sincerely appreciate your time in following and reading the results of this analysis. If you have any questions, opinions, or would like to share your thoughts, I warmly welcome and are open to receiving them.

To get in touch with me or to obtain more detailed information, please use the following means:

Email: anhvi09042002@gmail.com

LinkedIn: https://www.linkedin.com/in/anh-vi-pham/

Github: anhvi02 (Henry Pham) (github.com)

Kaggle: Anh Vi Pham (Henry) | Novice | Kaggle

I highly value all contributions and feedback and will endeavor to respond at the earliest opportunity. Thank you once again, and I look forward to hearing from you!