# Assignment - Telling a Story with Data
# Economics in the World of Big Data

- וידאו סיכום

https://drive.google.com/file/d/1thaT1Tj6B5l53-RSOqhnvKob4osBF6DT/view?usp=drive_link


Authors:

**Ariel Hedvat**

**Eitan Bakirov**

**Yuval Bakirov**

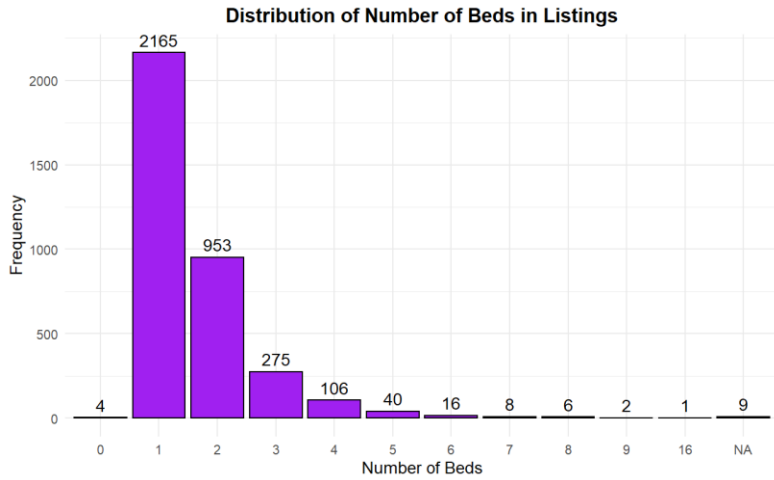

- Assumption regarding the available_category variable:
  For available_category = 0: The property was offered by the host but not ordered by a customer.
  For available_category = 1: The property was offered by the host and ordered by a customer.
  For a date that does not appear for a particular property, assume that the host did not offer their property for rent on that date.

- We note that in general it seems that each host offers the same amount of nights, in the region of 250 nights.

- Discount regarding prices: We assume that the price that appears in the listings_clean table is the initial nightly price that the host offered for his property when he first entered the platform, while the price that appears in the calendar_clean table is the nightly price that the host offered for that specific night and that this is the price actually paid by the guest.

## Part 1 - Descriptive statistics

### 1. Distribution of the number of beds–

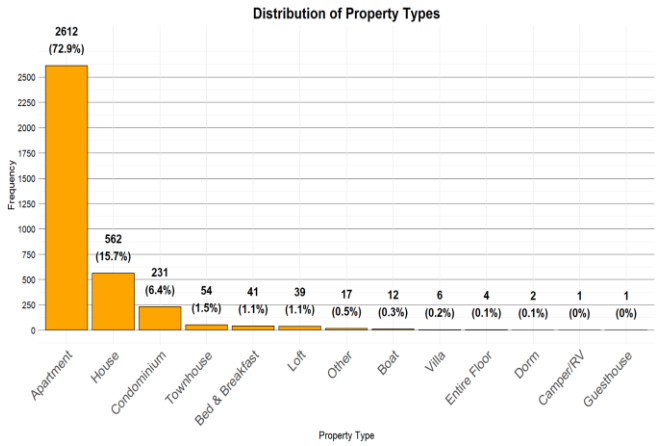



In the graph we will see the distribution of the number of beds in the property out of the properties of the city of Boston.

Most of the properties have a single bed. In fact, there are more entrants with 1 bed in the data than the number of properties with a different number of beds than 1.

Something suspicious that we will notice in the data is that a single bed can be a single bed and can be a double bed that is suitable for hosting 2 guests. There is no distinction in the data between types of beds. Such a figure can indicate the amount of guests who can stay at the property = affects the company's income.

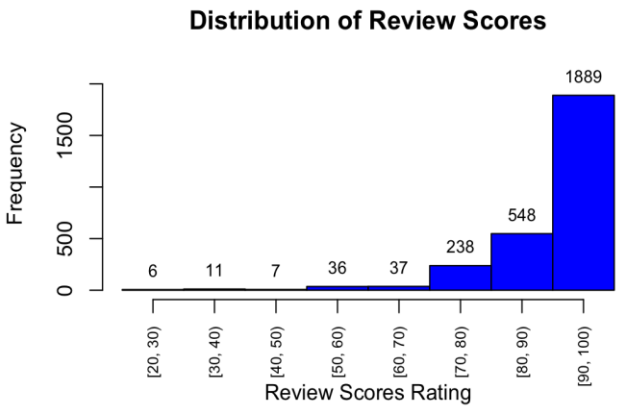
### 2. Distribution of asset types -–




In the graph you can see the distribution of assets according to the type of asset. Most of the listings in the data are apartments, with about 2612 apartments (72% of properties). This suggests that apartments are the most common type of property available on Airbnb in Data, which could possibly reflect the urban characteristics that characterize Boston. The presence of unique property types such as boats, dormitories, campers/caravans and guest houses does indicate diversity in the types of accommodation available, but the quantity can indicate that such properties are rarer. In addition, the high frequency of apartments may indicate a lack of diversity in the types of accommodation available, which can be a disadvantage for travelers looking for unique experiences.

3. **Histogram of property ratings -**

This graph shows the distribution of the average scores given to each property. You can see a high concentration of high scores, the majority between 90 and 100, and very few properties with low scores - indicating that most are highly rated by guests - indicating a positive overall hospitality experience.

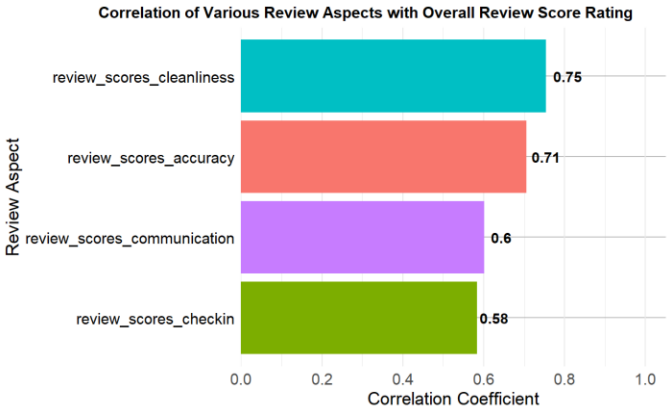



We suspect that there is someone hiding/distorting the grades for the better. It is likely that there will be quite a few properties with less good ratings, since Airbnb offers a wide variety of properties, with different prices, in different locations and with unique characteristics. Of course, looking ahead, we would like to continue to maintain a property system with high ratings. At the same time, understand what led to the low scores for some properties and what to do to improve the customer experience.

**<u>Graphs that identify interesting and important relationships between the variables</u>**

1. **Correlation graph between the review scores in different areas with the overall rating score-**

In this graph you can see the correlation between the review scores in different areas (cleanliness, accuracy of the apartment description, communication and check-in) with the property's general rating score. The higher the correlation, the stronger the correlation between the rating on a particular area and the general rating of the property.



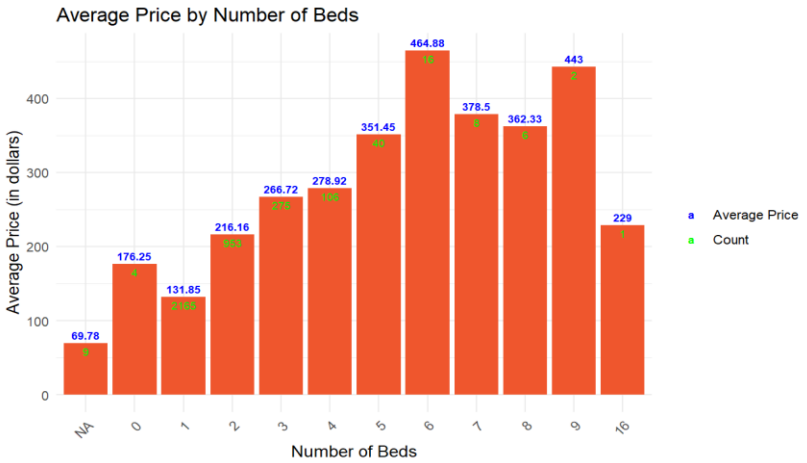

We wanted to look at areas that the host can have an influence on and thus in the future be able to improve the rating score of the property he offers for rent and succeed in attracting more customers.

The graph shows that, in general, there is a positive correlation higher than half for all parameters, so it can be concluded that if the host invests in his hospitality experience (cleaning the apartment, more precise explanations, good communication with the customers and a good check-in experience) he will have a positive effect on the rating of his property.

The graph shows that cleanliness has the biggest impact on the property's overall rating, suggesting that if hosts invest more in cleaning the property, it may have a positive effect on the property's overall score.

**2. Graph of average price per property per day according to the number of beds in the property -**

A graph of the average price per property per day by number of beds in the property, along with the number of properties for each number of beds in the property. There is a general trend of the average price increasing as the number of beds increases. This suggests that larger properties, as expected, command higher prices. The price does not increase linearly with the number of beds, which indicates that other factors also affect the pricing. There is a sharp drop in the availability of properties with more than 3 beds. Interestingly, properties with 0 beds (probably studio apartments) have a higher average price ($176.25) than properties with 1 bed ($131.85). The highest average price is for properties with 6 beds ($464.88), not the largest properties. This can indicate an optimal point for luxury rentals.

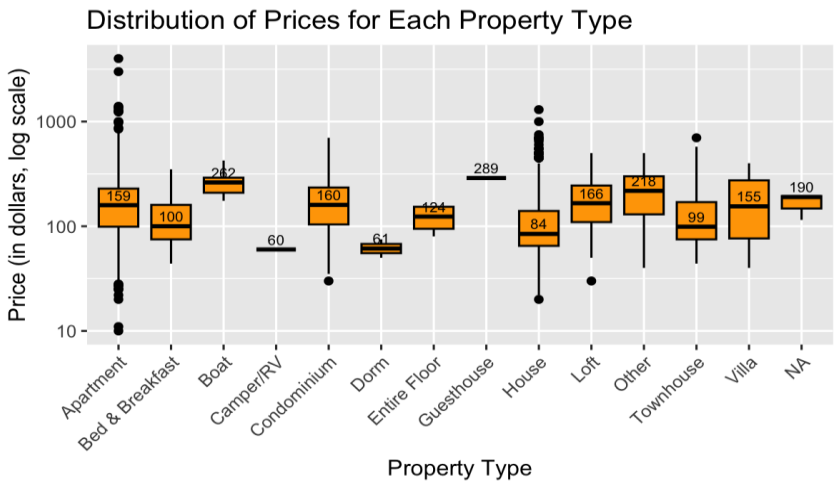



Overall, the number of beds in the property is not necessarily the main influence on the price of the property .

**3. The distribution of property prices by property type --**

We will see in the graph the price distribution of properties according to the type of property. We used a blog to deal with unusual values that make it difficult to display the data.

Property types such as apartments, houses, townhouses, villas and more show significant variation in their prices, this indicates that these types of properties can vary greatly in quality, size, location or services. There are several properties with low variance, but in these cases, this is due to their few observations in the data. Guest houses, boats, and "other" have higher median prices, suggesting they may be more luxurious or in higher demand. Dorms, cottages, camping, bed and breakfasts, and townhomes have lower median prices, possibly indicating more budget-friendly options. The presence of exceptions in most categories, especially for houses and apartments, can be due to extremely luxurious or very basic properties.



**Part2 – metrics (3 given)**

## 1. The growth dimension -

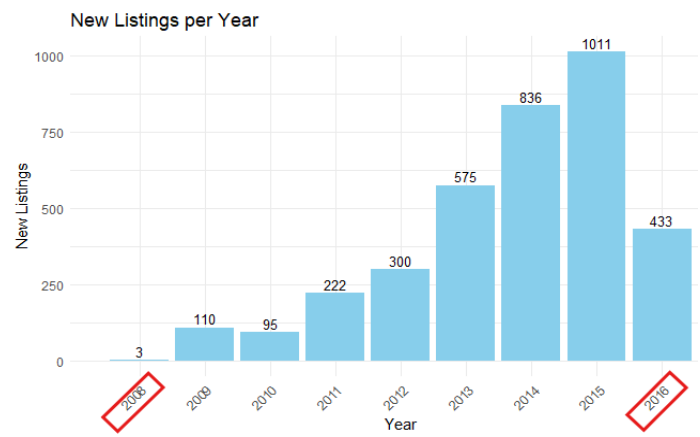
We have defined a metric that aims to see the annual growth of the property supply by counting the amount of properties that join the platform every year.

The formula is defined as follows: Annual Growth (New Listings) = New Listings in the current year

This metric captures the total number of new properties added each year, providing a clear picture of how the supply of properties on the platform is growing each year. However, this measure misses dimensions such as the quality of the rise of the assets on the platform. That is, whether the properties that joined the platform were ordered by the customers, and whether they manage to meet the market demand requirements.

By estimating the metric based on the data, it is possible to identify an increase in the number of new properties added to the system each year. Note that the year 2016 ends in September, so the data in the graph does not cover the entire year.

In the table of dates, the activity of the properties is recorded between Sept. 16 - Sept. 17 only! That is, it cannot be verified that the growth we identified in the number of properties offered each year in the property table, indeed brought with it more new customers each year (we lack data on property orders prior to 2016).




## 2. The price change dimension -

We defined a metric whose purpose is to show the average daily income on a property in a month. We chose to use a monthly average of the RevPAN index, which means Revenue Per Available Night across all given properties, it can be calculated using 2 methods:

### 1. RevPAN = ADR × Occupancy Rate per month

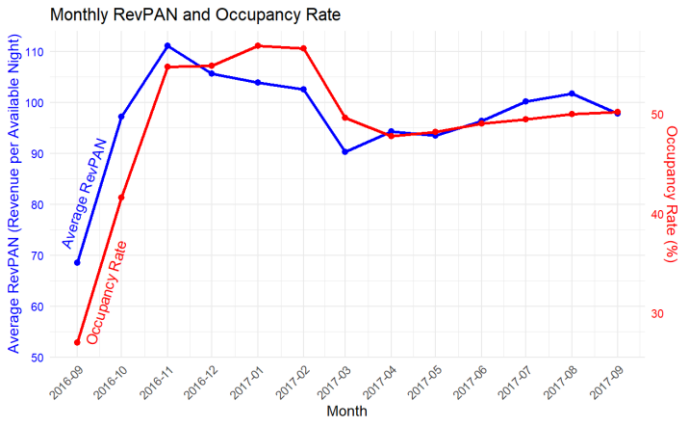
- ADR (Average Daily Rate): Total Listing Revenue / # Nights Booked
- Occupancy Rate: # Nights Booked / # All Nights per Listing
- Total Listing Revenue: Sum of price_dollars for a listing where available_category = 1
- # Nights Booked: Nights where the listing was rented out (available_category = 1)

- # All Nights per Listing: Both booked and unbooked nights within the month

**2. RevPAN = Total Listing Revenue / # All Nights per Listing**

This measure manages to capture the average income efficiency of properties by taking into account pricing and occupancy rates. Using an average RevPAN across all properties each month provides a snapshot of overall revenue efficiency for Airbnb to understand how to adjust prices appropriately. Outliers or poor performance can skew the average and hide specific problems. It also does not take into account external market effects on occupancy or prices directly.

1X1000 = **1000** -> RevPAN = 1000/1 = **1000** vs. 10X100 = **1000** -> RevPAN = 1000/10 = **100**




The RevPAN curve shows a decrease in RevPAN from the end of 2016 to the beginning of 2017, which indicates reduced revenue efficiency per available night during this period. However, there is a gradual recovery starting in mid-2017, indicating an improvement in pricing strategy and/or occupancy rates. In the context of price change intuition, we would like to change price in places where the RevPAN is low to increase the profit potential.

To do this, we will use the graph of the average occupancy rate per month to give a business recommendation and we will focus on areas where the RevPAN was low compared to the rest and check: for high occupancy rates, we would like to raise the price and for low occupancy rates, we would recommend lowering prices in order to attract visitors, increase the occupancy percentages as well as the total profit.

In the occupancy curve, we see a decrease in occupancy from the end of 2016 to the beginning of 2017, which aligns with the decrease in RevPAN during that period. As occupancy begins to stabilize and gradually increase from mid-2017 onwards, it supports the recovery trend in RevPAN.

This metric could be more accurate if we knew for sure the price actually paid without assuming that the price actually paid is the same as the price offered by the host.

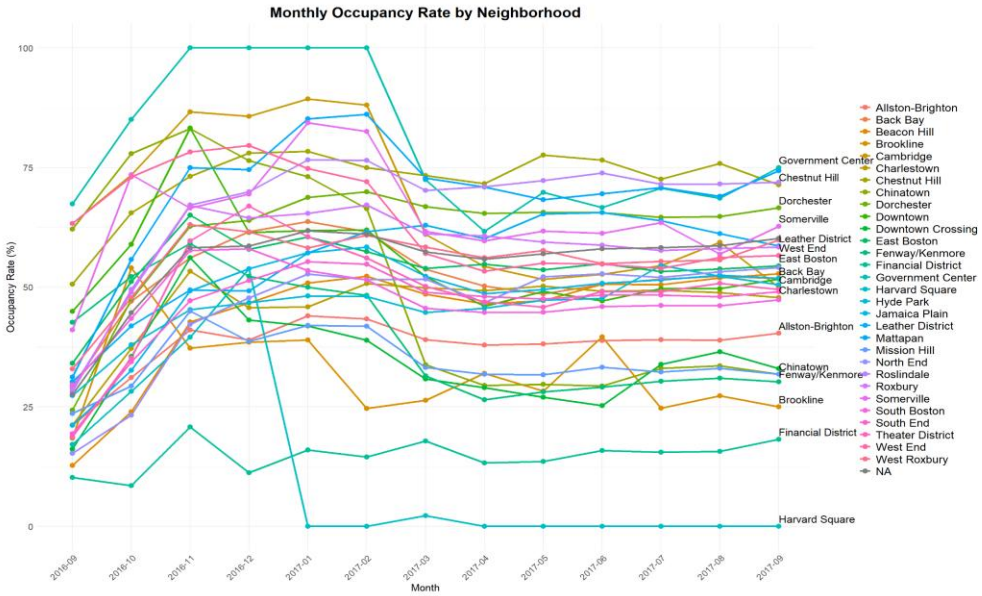
**3. The growth dimension by neighborhood -**

We defined the same metric as RevPAN, only this time we looked at the average income generated per free night within each neighborhood. This index, called Monthly Neighborhood RevPAN and is defined as follows:

**Monthly Neighborhood RevPAN = Total Revenue in Neighborhood / Total Available Nights in Neighborhood**

- Total Revenue in Neighborhood: סכום ההכנסות שנוצרו מנכסים בשכונה מסוימת עבור חודש מסוים.
- Total Available Nights in Neighborhood: המספר הכולל של הלילות הזמינים ע"י המארחים בשכונה מסוימת עבור חודש מסוים.

The index captures the economic performance and growth potential of different neighborhoods, and thus can provide insights into neighborhoods that earn more and less, and accordingly propose a future economic strategy. However, it does not refer to the size of each neighborhood, the quality of the neighborhood and how touristy it is, information that is missing in the data we have and may help us understand how realistic the growth of the neighborhood is given the reality. The graph shown indicates variation in RevPAN between neighborhoods, with some areas consistently outperforming others and those trending downward.



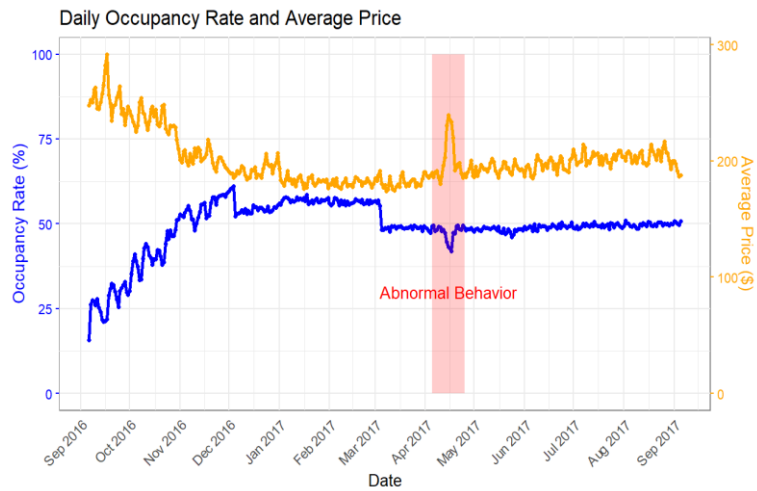

We will look at the number of properties there are (as of September 2016 - lack of information regarding properties that joined later, something that could clarify the understanding of the business situation) in each neighborhood and based on that we will understand the state of our business. It is evident in the graph that there are neighborhoods with few properties and low RevPAN, and neighborhoods with many properties and high RevPAN. We would like it to be the opposite to increase Airbnb's profits, so that we could be better off.

We would suggest that Airbnb check which neighborhoods have low growth potential and understand whether it is a lack of supply, a disproportion between occupancy percentages and price (also here it is useful to look at the occupancy percentages in the neighborhoods and accordingly propose a lowering or raising of prices), neighborhoods that are considered least good or less touristy and then accordingly propose Property owners have to lower prices to attract visitors. In addition, Airbnb should focus on improving supply in high-performing neighborhoods and maintain growth in these neighborhoods, and on the other hand, develop strategies to improve performance in low-performing neighborhoods to increase growth there.

## Part3 – outlier detection

### Time range with unusual behavior -

Let's look at the graph containing the Occupancy Rate (discussed in part 2 - price change dimension) and the average price each day:

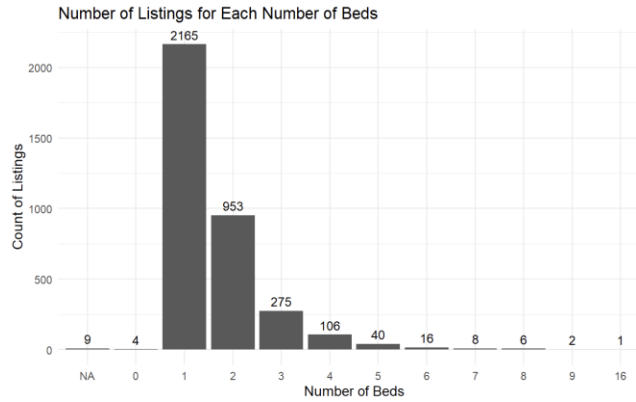



We will note that in the period 12/04/2017 to 19/04/2017 there is a decrease in the occupancy rate and an increase in the average prices.

A short research we did found that the Boston Marathon took place on these dates. In our opinion, due to the marathon planned for April and the growing demand for inventions in the area, the hosts raised prices drastically, which ultimately caused a decrease in actual orders. These are real data, and their omission will harm the continuity and uniformity of the data. We would like to mark these abnormal values, to know how to draw conclusions from them for the rest. A way of coping that we will propose is to add a Boolean variable 'Event', which will signal a period that may have an abnormal effect on the data.

### Number of beds-

We will note that there are some unusual values: properties with 16 beds, 0 beds and NA.



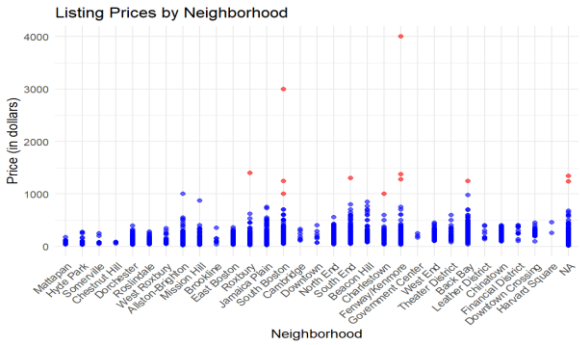

According to the description of the room with 16 beds, it can be understood that it is indeed a large room and according to the definition of the host, 16 people/beds can be accommodated. We will leave the figure so as not to harm the diversity of the data.

According to the description of the rooms for the properties with 0 beds - all are studio apartments. Studio rooms usually contain one bed, so we will convert the number of beds to 1.

There are 9 properties with an unknown number of beds (NA). From a brief investigation, although some of them are double listings and without reviews, they were indeed rented throughout the period, so we will convert their number of beds to the average of the data, that is, one bed.

**price-**

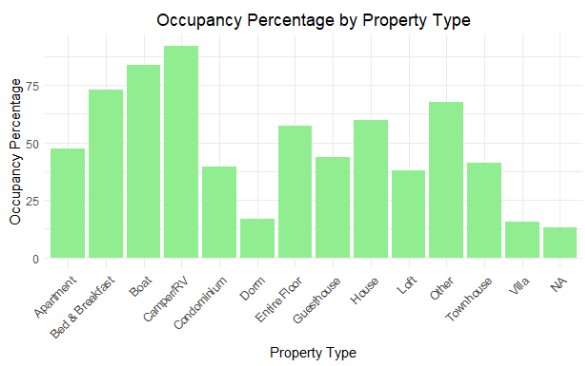
From a horizontal look at the prices in each neighborhood, it can be seen that most of them range in price between 0 and 1000, with some unusual prices such as 3000 and 4000. There are 12 properties worth 1000 or more, 10 of which are apartments, and 2 houses. 8 of these 12 properties were ordered - meaning there was demand for most of these properties despite the high price.

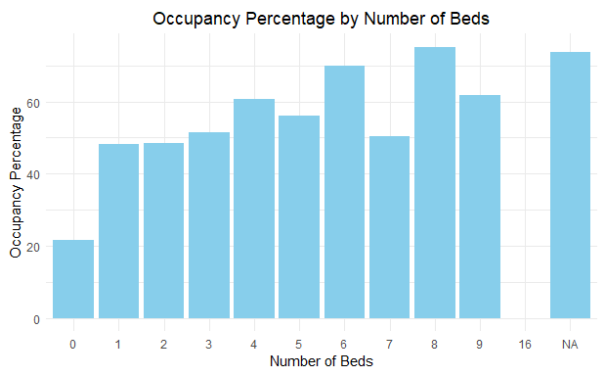



The abnormal values of these prices can be due to several reasons that are difficult to understand explicitly from the data. Looking at these properties you can see that they are all in central neighborhoods. In terms of handling these values, we will not touch so as not to harm the diversity of the data.

**Part 4 - Business recommendation:**

Airbnb brings money to the company by taking ~3% of the total nightly price that the host charges. We would recommend the company to create incentives through benefits (for example, incentivize hosts to become a superhost and add a benefit also for the guests who leave scores that help us create a more accurate analysis for the property) , campaigns and marketing campaigns to increase the number of properties offered in each neighborhood and especially neighborhoods with high profit potential or neighborhoods that received a high satisfaction score (such as the Financial District). So, by and large, the goal is to create an increase in both supply and demand that will result in an increase in rentals and thus an increase in the host's income.




According to the analysis we performed, we saw that there is no diversity in the offer of the properties offered, we would like to change this.

It seems that the supply of properties with 1 bed is the highest, but actually the occupancy percentages were higher for properties with a larger number of beds (this is also less good because we saw that on average a property with 1 bed is the cheapest). In addition, it seems that the supply of apartment properties is the highest but only with 50% occupancy, therefore we would suggest perhaps increasing the supply in properties with higher occupancy percentages (such as boats and houses) and see if there is a change. This is in order to increase diversity to reach different populations.

In addition, we would like each host to reach the maximum profit potential that it can reach (we can tell this by looking at the RevPAN combined with its occupancy percentages) and cause quality price changes. Therefore, we recommend that Airbnb have such a view that can alert the host based on previous data whether it is worth raising or lowering prices in order to create the perfect balance between the price per night and the number of visitors.

From what we can see at the moment, we would suggest that Airbnb generally recommend to hosts next year to slightly lower the price per night in March. If the occupancy rates increase and the average income increases, we will advise them to continue with these prices in the following year as well.