Discovering Trends in Rental Airbnb Over Time

Group I

Ravish Kamath, Yilin Chen, Vi Nguyen

MATH 4939 M- Statistical Data Analysis using SAS and R
Winter 2022-2023
Mathematics & Statistics
York University
Canada
April 10, 2023

1 Business Problem

Business Context: In 2007, recent college graduates Brian Chesky and Joe Gabbia embarked on a mission to solve a distinctly Millennial problem: affording rent in an expensive city. They founded a company named AirBed & Breakfast. Although initially it was just an air mattress in their living room with breakfast service, it soon evolved into thousands of listings and tens of thousands of users. Soon enough, AirBed & Breakfast (Airbnb) extended its reach across the globe, ballooning to a valuation of over $30 billion in March 2017. Airbnb has been a great boon to renters on their platform, solving for them the same problem that its founders had a decade ago, affording their apartment rent via subletting it out to others when they weren't using it. However, it has forced hoteling industry incumbents to find new ways of differentiating their services, as the increased competition has driven down prices and put a sizeable dent in the low-to middle-end business.

Business Problem: What trends can you find in the Airbnb rental calendar over time, and how might these be explained by listing-specific and/or neighbourhood-level factors?

Insights into this problem: The company would be able to identify trends in demand and price by keeping track of the rental calendar data over a period and developing effective strategies to enhance its operations. For instance, the company could decide to invest in amenities that are well-known and liked by customers, and modify the price based on seasonal demand. Moreover, the company would acquire a deeper knowledge of the variables that influence the demand and price for short-term rentals in certain neighbourhoods by merging Airbnb calendar data with demographics and income data.

2 Business Impact

Analysing US Airbnb data and combing them with other relevant data sources has significant business impacts. Here are some of them:

1. Identify demand and pricing trends: We can determine the trends in demand and price patterns for regions and times of the year by analysing Airbnb listings and calendar data. Pricing strategies and marketing initiatives can benefit from this.

2. Optimize marketing, customer targeting and making better investing decisions: Using Airbnb data in conjunction with demographics and income info can help the company better understand their targeted customers and locate the markets with significant growth potential. We may discover regions with high housing costs by merging Airbnb data with economic and housing-related data at the state and zip code level. This will help make marketing plans and outreach initiatives.

3. Competitor analysis: Data analysis on Airbnb may also provide information on pricing strategies, client targeting, and areas of emphasis for rivals, helping companies position themselves more effectively in the market.

To summarize, the company is optimizing its operations, and find development prospects to benefit greatly through the rental data analysis

3 Data

In our analysis we will be focus solely on **calendar.csv** and **listings.csv** dataset.

Calendar

This dataset contains 21,941,235 observations and 5 columns. Those 5 columns are:

1. listing id (int): ID of the listed property.
2. price (float): The price of a one-night stay, in U.S. dollars.
3. date (string): Date for which data is available. Format yyyy-mm-dd.
4. metro area (string): The metropolitan area, the listed property is located in.

The large number of rows is due to the dates listed. It covers everyday from the year of 2016 till the year of 2018. Furthermore the metropolitans covered in this dataset consist New York, DC, Chicago, Boston and Denver. When we took a look at any NAs or errors, we found that only the price variable had roughly 60% of the data missing. We will handle this issue in the Data Processing section of our analysis. This data will be very useful because we can take a look at trends over a period of time, such as by year, day or month. We can see the fluctuation of prices, as well as how different metropolitan areas affect price.

Listing

This dataset contains 59,824 data points with 29 columns. We won't list out each one in this report, however we will show which columns we decided to keep for our analysis.

1. accommodates (string): Number of customers the property can fit
2. availability 30 (int): Number of nights the property is available in the following month
3. bed type (string): Type of beds available
4. bedrooms (int): Number of beds in the property
5. city (string): Neighbourhood/city the property located in
6. id(int): ID of the listed property
7. instant bookable (string): Whether the property allows for instant booking
8. latitude (float): Latitude of property
9. longitude (float): Longitude of property
10. metropolitan (string): Metropolitan area the property is located in
11. price(float): One-night rental price of the property, in U.S. dollars
12. property type (string): Type of property
13. review scores location (int): Customer review of the property's location
14. review scores rating (int): Overall customer review of the property's rating
15. review scores value (int): Customer review of the rental value
16. room type (string): Type of the rooms available
17. state (string): State the property located in
18. zipcode (int): Zip code of the property's location

There is quite a few things to fix in this dataset. We have a bunch of NAs in various columns and rows. We want to fix this so that we can use it as a part of our analysis. Again this will all be covered in our Data Processing section of our analysis. Overall this will be our main info dataset of all the rental properties. We can use this dataset to help describe the trends we found in the calendar dataset.

4 Methods

4.1 Data Processing

Listing

The first thing we did was remove columns that we personally believed will not be useful for our analysis. When it came to dealing with NAs we performed a couple of methods:

- Replace NAs with unknowns: accommodates, bedrooms and city.
- Replace NAs with the median: review scores location, value and rating.
- Drop Rows that had NAs in zipcode since it was roughly 1% of our data.

Calendar

For the calendar dataset the biggest thing was replacing the prices. Since we had the prices of all the listings from our listing dataset, all we did was replace the NAs with those values, where we matched it based on the correct ids. That took care of the 60% price data we were missing. Also, we had some outlier listing ids that did not match with the listings id, which were roughly 14 of them. We removed these 14 data points since we already have a lot of data points available for our analysis, without worrying about losing vital information. Finally, we converted our date column into a date-time object so that we can do further analysis in terms of months, days and/or years.
After all this cleaning, we decided to merge both data files using the .merge function. We merged based on the ids of both datasets. This new dataset is called `listings_calendar`.

Listings Calendar

We still need to clean this dataset a little bit more.

- Convert all the ids into type string
- Replaced the calendar prices values with the prices from the listing data set. We tend removed the listing prices column.
- Replace the 't' or 'f' to Yes/No in the `available` and `instant_bookable` columns.
- Removed `metro_area` as we have the metropolitan data from the listings dataset.
- For the `metropolitan` data, we made it all upper case to be consistent.

4.2 Visualization

Overall Picture

Let us first take a look at the overall price distribution for all the data points, and which year is most of our data coming from. From Figure 1 we can see that most of prices are distributed around $0 - $500, and outliers past that. In terms of the year that has most of the data, according to Figure 2, we can see that 2017 will have most of the info data.

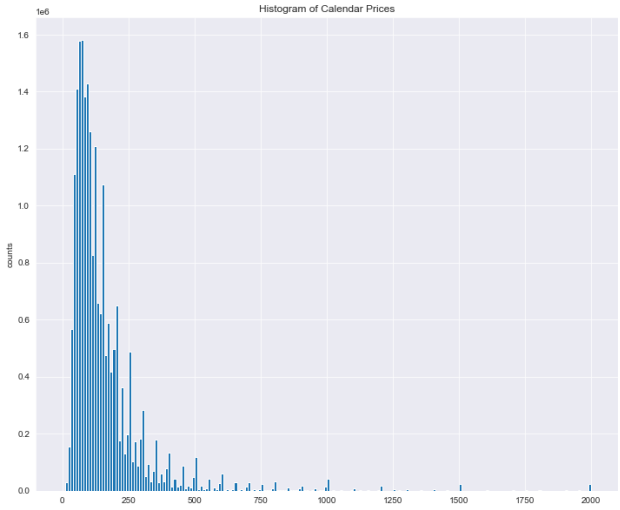


Figure 1: Histogram of Prices

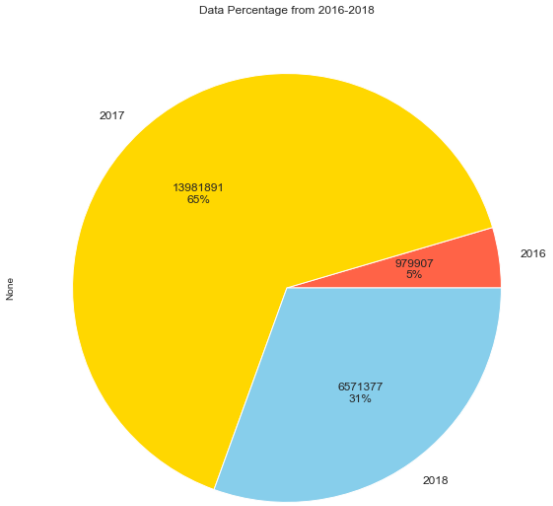
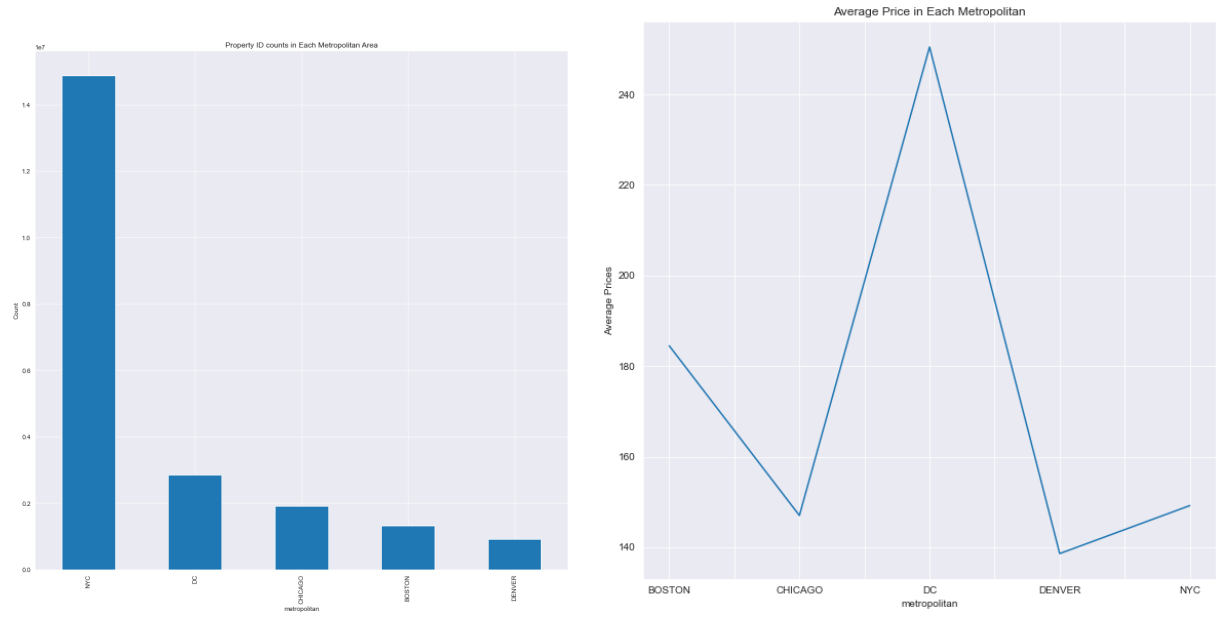


Figure 2: Data Percentage from 2016-2018



(a) Property ID counts in Each Metropolitan Area

(b) Data Percentage from 2016-2018

Figure 3

Since we want to focus on trends that can be explained by neighbourhood levels , it seems only fair to see how our data is distributed between the major metropolitan areas. As shown in Figure 3 (a), we can see that most of our data is coming from New York Metropolitan area. Furthermore, in Figure 3 (b), We can see the average rental prices in each metropolitan. It is quite surprising to see that DC outstrips the rest of the metropolitans (especially NYC) in terms of rental prices.

Figure 4: Monthly Prices In Each Metropolitan Area

Now that we have seen the average prices for each metropolitan, we can also make a visual to see any initial trends of each metropolitan over time. We tried to see any trends in years and days of all the metropolitans, however it seemed very stable. However, when looking at the months, we saw some trends that could be interesting for our analysis. According to Figure 4, we can see some sort of a cyclical trend, which is quite evident in Chicago. It seems like during the summer months (May - August) we see that the prices increase while it goes down during the fall/winter period.

Exploring Most Expensive Neighbourhoods

An important question to ask is justifying the prices in expensive neighbourhoods in each metropolitan. We may be able to see trends across all the different expensive neighbourhoods and factors that influence price. Let us see if we can find the most expensive neighbourhoods in each metropolitan. According to Figure 5 below we can see a list of the most expensive neighbourhoods.

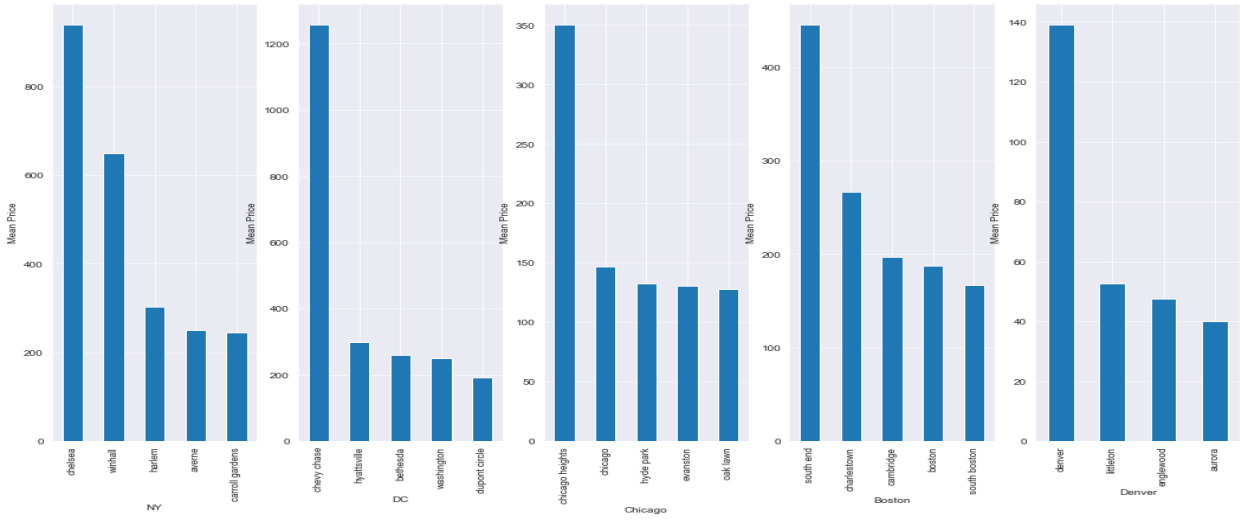


Figure 5: Top 5 Most Expensive Neighbourhoods

Some very interesting neighbourhood results show up. We google some of these areas to find out more info about why these places are some of the most expensive Airbnb rentals in our dataset.

- **New York: Chelsea**
  This is quite interesting because we would assume that New York neighbourhood area would have the most expensive Airbnb, however this is not true. This neighbourhood could be for the richer folks, or water facing, or other features that can increase the price. When looking on Google about this area, we can see that it contains luxury high-rises and trendy attractions like High Line. Furthermore, this is one of the best places to live in New York according to several websites.

- **DC: Chevy Chase**
  Chevy Chase is a neighbourhood in the north-west of Washington. This is also quite surprising because this is not in the heart of Washington where many people visit for tourism etc. The key features of this area, based on Google, is that it has a strong sense of community, the lifestyle is quite laid back compared to the city, and features more single family homes rather than condos. This can be of great interest as to why a suburban neighbourhood area would be more expensive than renting an Airbnb in the heart of Washington.

- **Chicago: Chicago Heights (also known as Heights)**
  Heights is a suburb of Chicago, about 30 miles south of down town. To me, this is also another interesting feature, since it is not in the heart of Chicago city. One key feature about this neighbourhood is that it is not the safest community in America in general (according to Google). This is quite unexpected, why people would want to pay higher in such area. There may be some discrepancies with this data.
- **Boston: South End**
  South End is a neighbourhood in the heart of Boston. It is distinguished from other neighbourhoods by its Victorian-style houses and many parks around the area. It has many features that well explain why rent in this area is quite high.
- **Denver: Denver**
  As expected, Denver (which is the only major city in Colorado) has the highest rental prices in Denver.

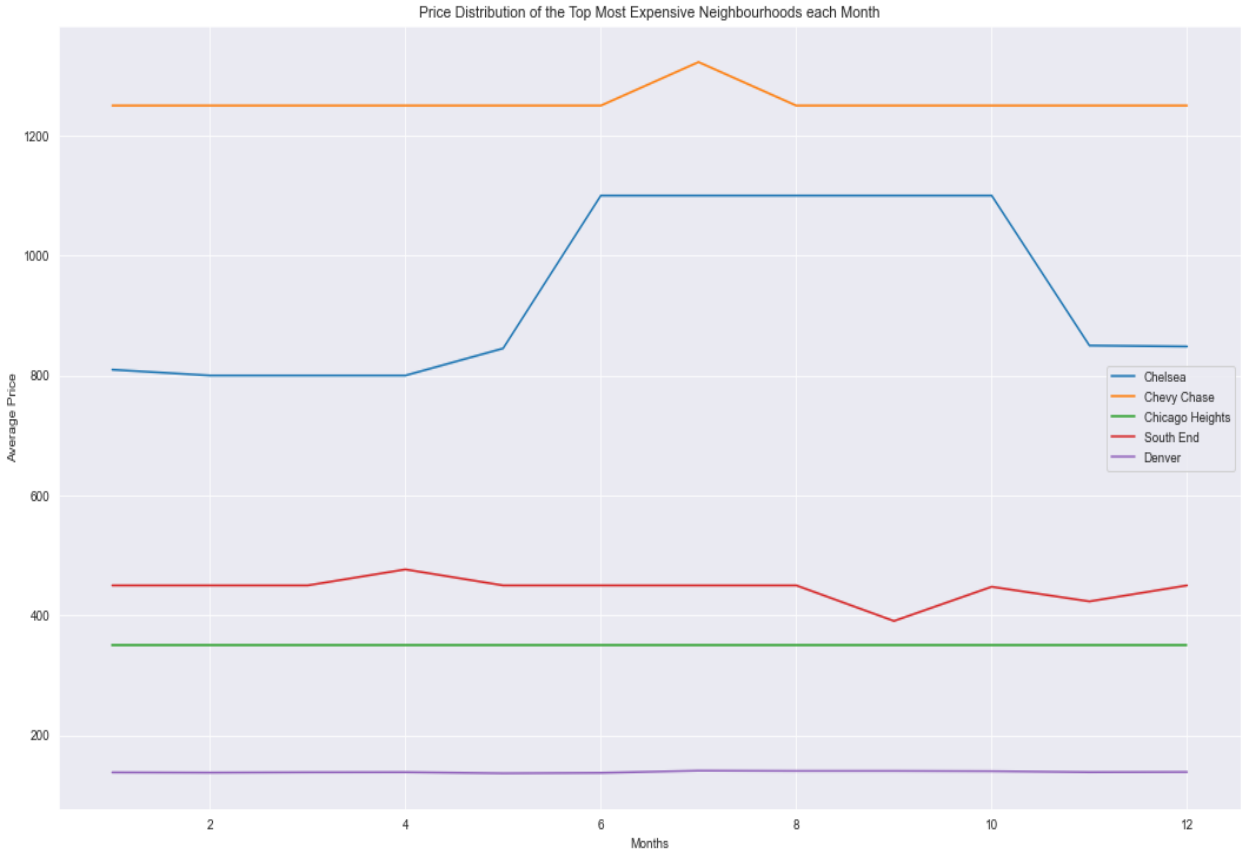


Figure 6: Monthly Average Prices of Most Expensive Neighbourhoods

We then tried to take a look at the trends of average prices in each month for these expensive neighbourhoods. According to Figure 6 we can see that only in Chelsea, NY do we see a huge change in prices during the summer, other than that, we don't see much differences in price. We also wanted to look at the availability for these properties. Does the demand for these rental properties justify their prices? Well, according to the Figure 7 we can see that almost every neighbourhood except for Denver are always available, which tells us that these properties don't have much demand. In the case of Denver there is almost an equal amount of demand as availability. Finally we would like to see what is the most popular property types in these neighbourhoods. Depending on whether the neighbourhood is a borough, city or a suburb, houses and apartments are the most popular. For example in Chelsea New York, there are many luxury apartments,

hence why we would see apartments as the most popular choice of rental, versus in Chevy Chase its more suburban hence houses are the most popular.

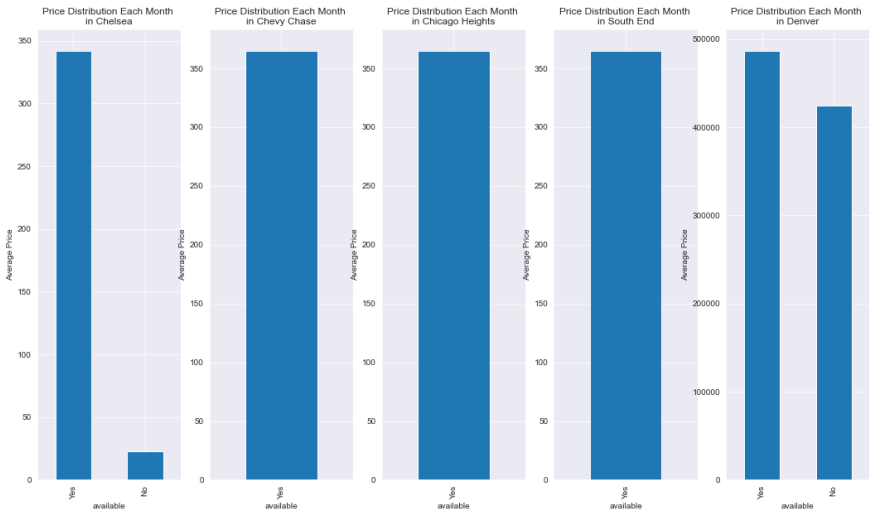


Figure 7: Availability in Expensive Neighbourhoods

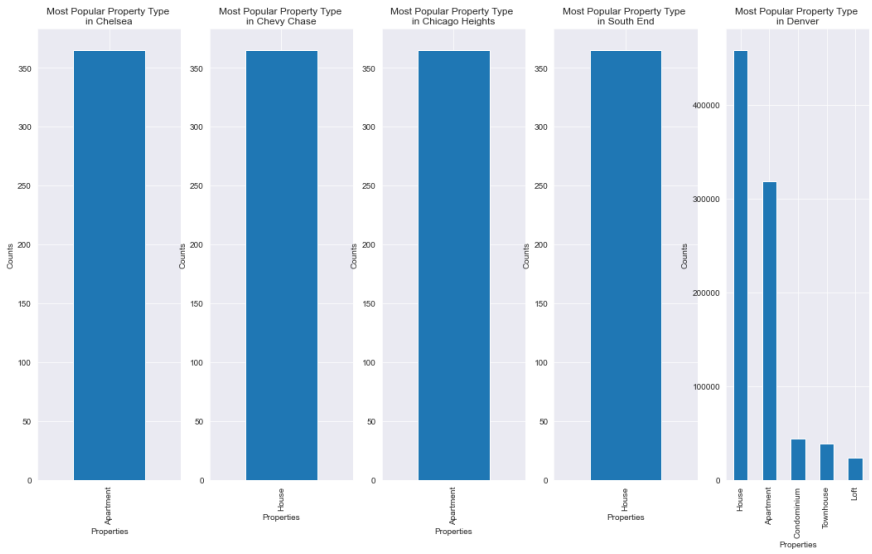


Figure 8: Most Popular Property Type

Exploring Most Popular Neighbourhoods

We have looked at some trends in the most expensive neighbourhoods, now we can also look at some trends across the most popular neighbourhoods in each metropolitan area. As we can see in Figure 9 clearly the city names of the metropolitan areas are the most popular spots. Within this cities we tried to take a careful look on how the prices fluctuate throughout the months of the year. In Figure 10, we can see that for almost all of the cities, there is a spike during the summer period vs. the rest of the months of the years. Those particular months are from May till August/September. We see a big price decrease after the summer period for most of the places except for New York, which seems to only fall during the beginning of the new year until Spring. Based on this we can see that New York starts low, but steadily rises throughout the year. What is also interesting is that we can visual the most expensive cities and how they relate to each other. Clearly, in this graph on a month to month basis, Chicago is the most expensive popular city, while New York

and Boston are similar in average price. Then Washington and Denver are within the same price bracket range. This does show trends on the pricing may work in these popular areas. Cities like Chicago, New York and Boston are more metropolitan compared to Denver and Washington which can be more suburban/ less skyscrapers type. This may have a good indication on the trend of listing prices as well as the property types over time.

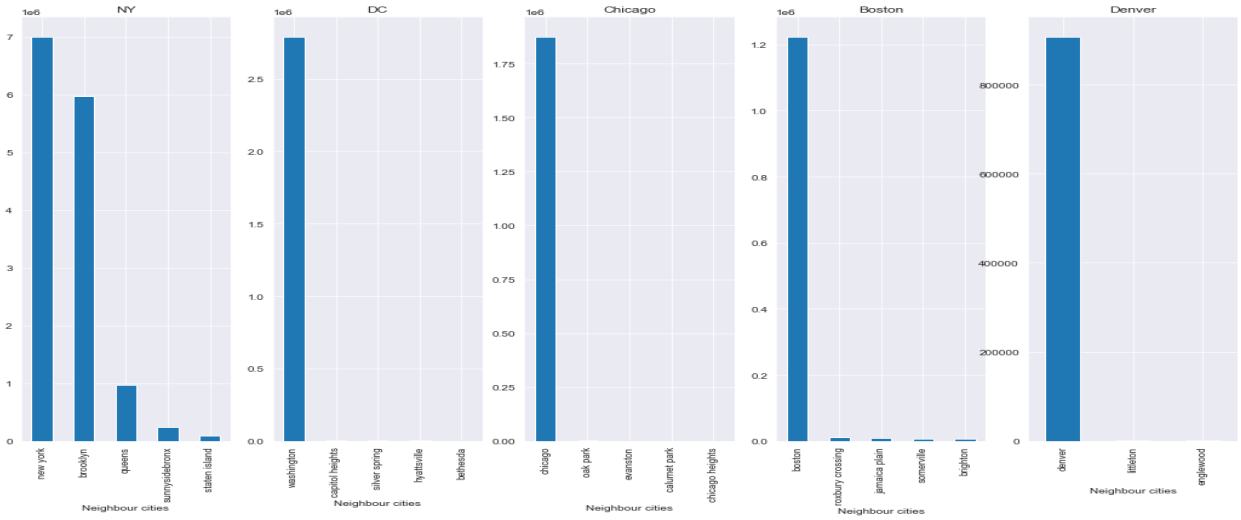


Figure 9: Top 5 Most Popular Neighbourhoods

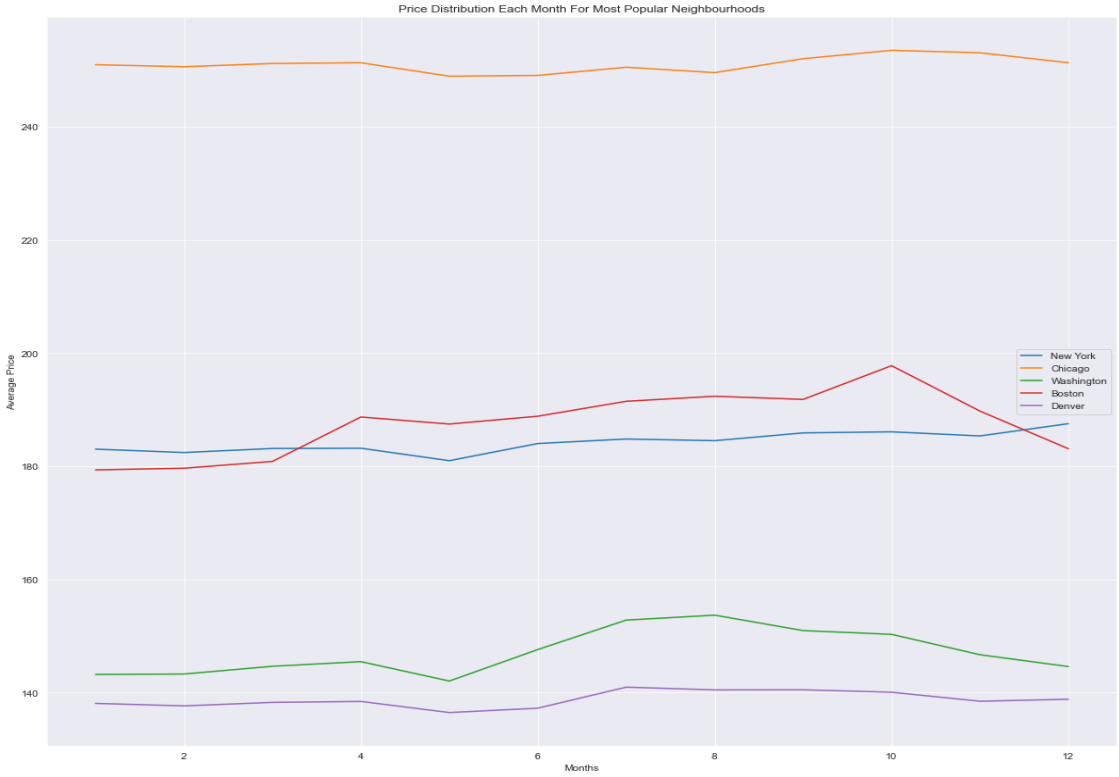


Figure 10: Monthly Prices of Each Popular Neighbourhood

Because of the trends we saw during the summer for all the popular neighbourhoods, we decided to focus on those specific months, and see any trends of the property types. In Figure 11, we can clearly see that apartments are the most popular during the summer except for Denver which would be houses. This is a definite trend we see throughout all the neighbourhoods which makes sense since these cities are high in terms of population density. So apartments would be the most available and most popular choice of rental property type. Since apartments are the most popular during the summer, we wanted to see how are they priced during each day of the week. In Figure 12 we can see that Friday and Saturday are where the prices increases for popular city, and then falls back down on Sunday. This may be due to the fact that people tend to visit these cities on the weekend and not during work time.

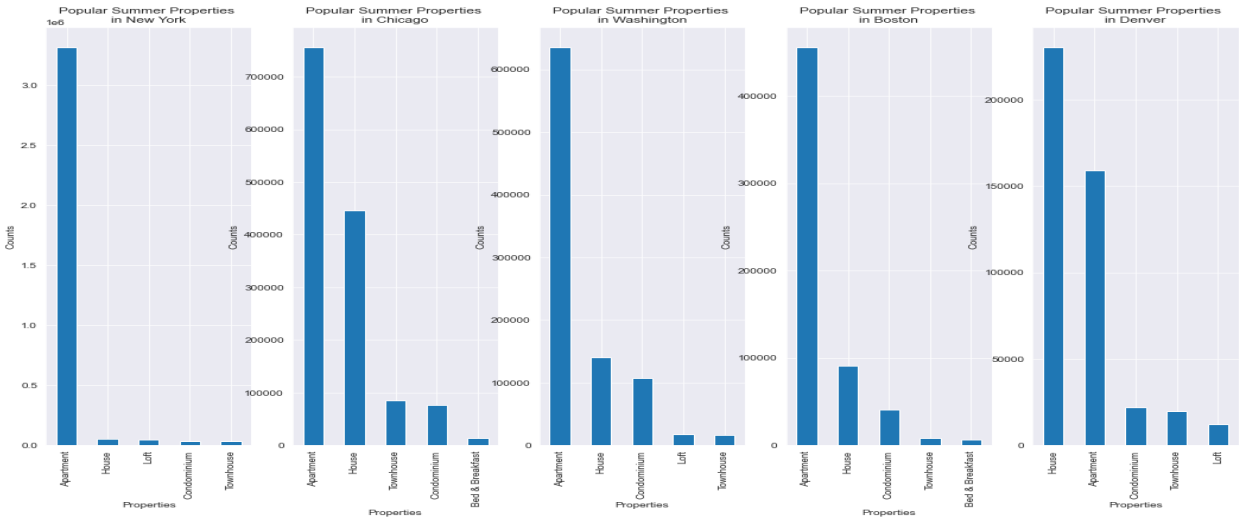


Figure 11: Top 5 Most Popular Neighbourhoods

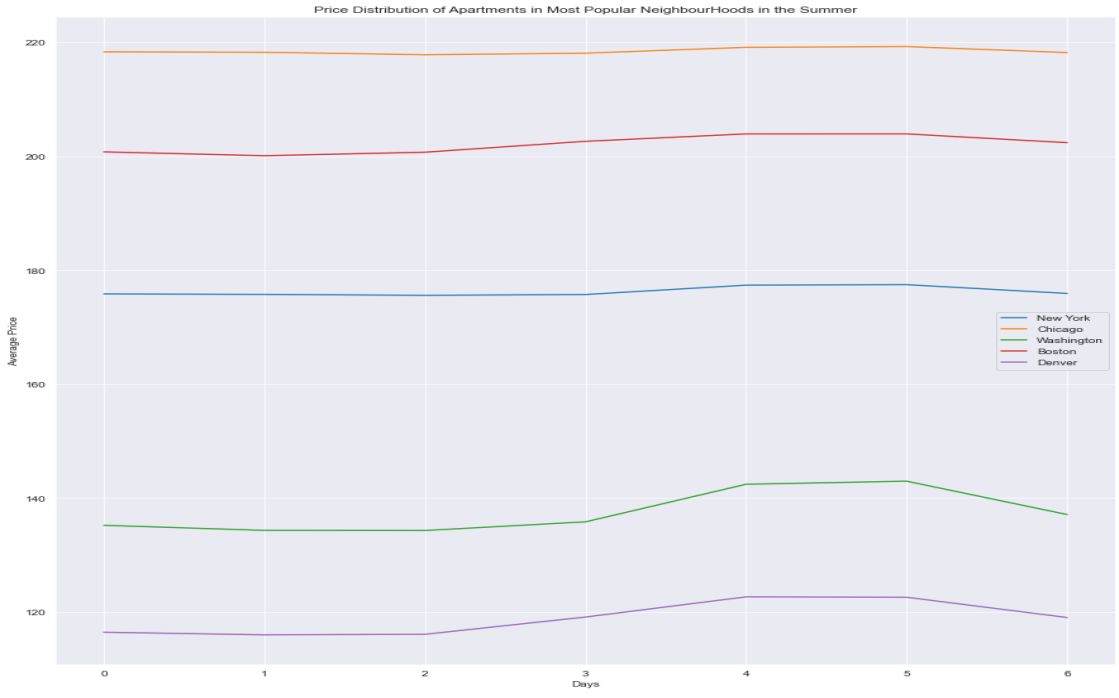


Figure 12: Monthly Prices of Each Popular Neighbourhood

4.3 Model

Now that we have taken a good look at what trends we can find in our calendar dataset merged with listings, it is time to see if we can build a model, using the previous knowledge of trends we found. Unfortunately, we cannot use the listing calendar dataset as there are too many data points, and we believe that the listing dataset will provide more than enough information for us to build a prediction model.

4.3.1 Multiple Regression

For our analysis, we firmly believe that the most simple model is the only model required for the listings dataset. When it comes to regression modelling, it is very important to discuss some assumptions we believe are true, for us to conduct this model.

- errors are independent and identically distributed with mean 0 and variance is σ^2 which does not depend on any of our covariates we define.
- The probability distribution of the errors is Normal.

One key issue was to make sure that our response variable (which will be the price) is normally distributed. Recall from Figure 1, the histogram indicated that prices were really skewed to the right. Hence we can apply a log transformation, to make it more normally distributed as shown in Figure 13. Finally, the variables we use for our model, that we believe are the best predictors to explain the prices are as follows:

- accommodates
- bed type
- bedrooms
- city
- instant bookable
- metropolitan
- property type
- review scores location
- review scores value
- room type

Most of these covariates are categorical, while others were type object which we converted into an integer primitive type variable (bedrooms). With these covariates our prediction model is:

$$\textbf{Price} = \beta_0 + \beta_1 \times \textbf{accommodates} + \beta_2 \times \textbf{bedtype} + \cdots + \beta_9 \times \textbf{roomtype} \tag{1}$$

By applying equation (1) to our listing dataset we get the summary statistic as indicated in Figure 14. With an $R^2 \approx 0.6$, we can see that our model does do well from a statistical test point of view. This value essentially tells us that 60% of the variation is explained by the covariates we have chosen. Of course we can add more factors to increase our R^2 value, however it is important to have a simple model in order to better understand how the variance is explained by your predictor variables.

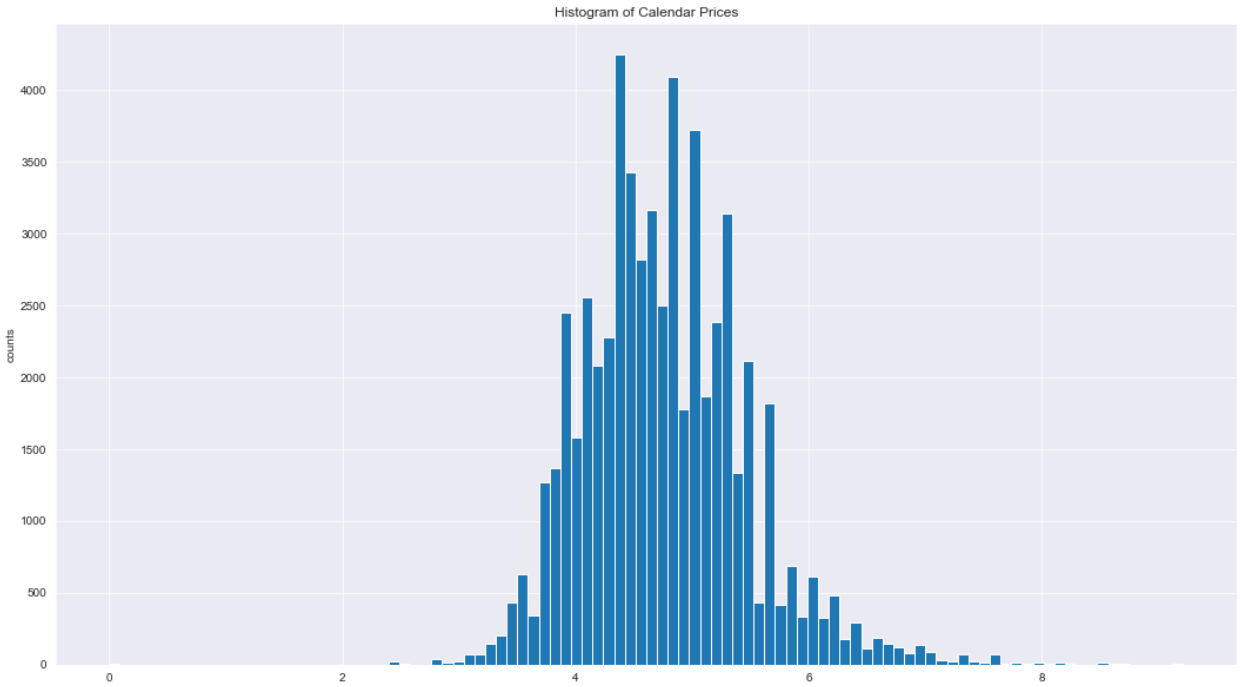


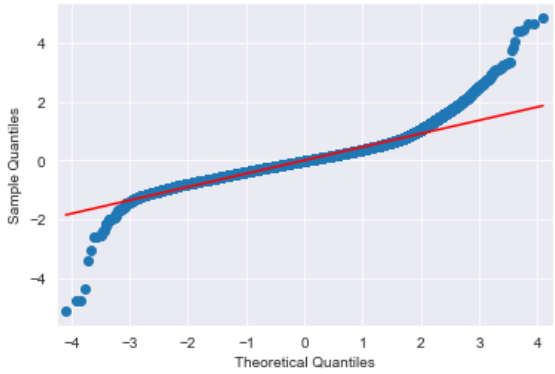
Figure 13: Log Transformation on Price

OLS Regression Results			
Dep. Variable:	price_log	R-squared:	0.592
Model:	OLS	Adj. R-squared:	0.590
Method:	Least Squares	F-statistic:	341.2
Date:	Sun, 09 Apr 2023	Prob (F-statistic):	0.00
Time:	23:18:39	Log-Likelihood:	-29563.
No. Observations:	47195	AIC:	5.953e+04
Df Residuals:	46994	BIC:	6.129e+04
Df Model:	200		
Covariance Type:	nonrobust		

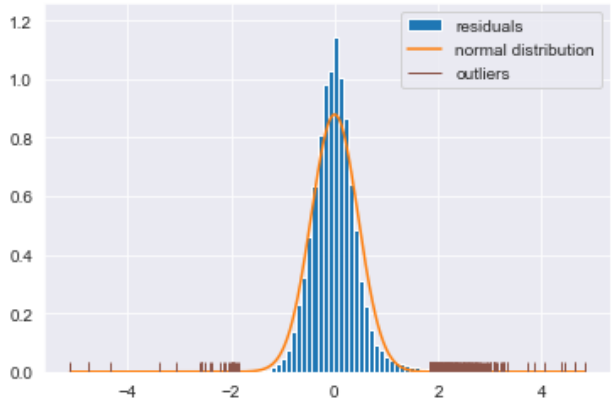
Figure 14: Model Summary

We also tried to look at which covariates were considered statistically significant by checking their p-values: Based on the model summary results indicate that :

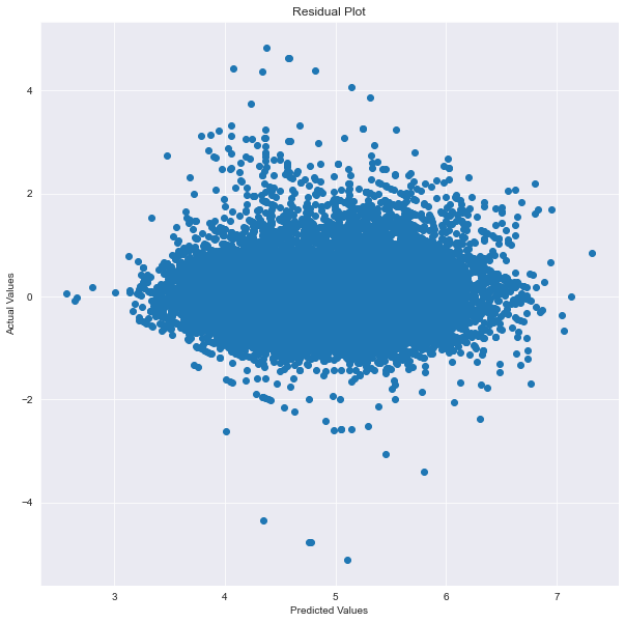
- availability 30
- room type
- accommodates
- bedrooms
- property type



(a) QQ Plot



(b) Residual Plot Checking For Normality



(c) Residual vs. Predicted Values

Figure 15

- city
- review scores location

are considered to be statistically significant covariates with a critical value for rejection $\alpha = 0.05$. Finally, it is not enough to just build the model, but to also test how well our assumptions hold. In Figure 15 we several plots that perform observational tests on the residuals and normality assumptions. As we can see all the plots indication good signs that the residuals do not get out of hand, and the QQ plot indicates that our errors are normal and linear. Overall, these are very good signs for a model prediction.

To summarize, we built a multiple regression model with an $R^2 \approx 0.6$ to see whether we can use the factors that we saw in our EDA analysis to help explain the variation of the prices in the listing dataset. If we had the time, we would have definitely done some Time Series analysis on the calendar data set, to create a forecasting model. This would be very useful for future renters to use to identify the right property for their budget!

5 Concern

While analysing US Airbnb rentals data is able to provide valuable insights, it also sheds light on some limitations and concerns:

- Data quality and sampling issues: If the sample we collected was not sufficiently random or was self-reported, the findings we get may be biased (response bias and selection bias) and may not be representative of the total population. The accuracy of the results is influenced by the quantity and quality of the data gathered.
- Privacy issues: The use of personal information, such as age, salary, and other details, leads to privacy problems. While using these data for analysis, the company must be extremely cautious. The company should also ensure that they are utilising the data in line with regulations and laws to prevent information leaking.
- Possible issues when interpreting the results: The business must take precautions when interpreting the findings, especially correlations. Correlations do not mean causation, and confounding variables might cause inconsistent results by introducing variables that are outside the purview of the analysis.

6 Conclusion

Recall that our business question was to identify trends in the Airbnb rental calendar over time, and how these might be explained by listing-specific and/or neighbourhood-level factors. The business impact of this question will allow managers and decision makers to utilize trends to drive up future business and also identify ways to make customers more happy with the use of Airbnb.

Our first step was to process the data in such a way that we do not remove valuable information, but also not be mislead by incorrect data inputs. We decided to merge the listing and calendar datasets to proceed with our analysis.

Our focus was to discover useful trends in neighbourhoods level factors. By looking at the most expensive neighbourhoods, we saw that price was not driven by demand but more by the location and property type. We saw that it was a 50/50 on apartments and houses for the popular property in the expensive locations. When it came to popular neighbourhoods, we noticed that prices were driven higher during the summers, and specifically during the Friday and Saturday of each week. Apartments were the main property types in most of the popular neighbourhoods due to being located in massive metropolitan areas.

Finally, we developed a prediction model using a multiple regression model with an outcome of $R^2 \approx 0.6$. Our assumptions were not violated at all, and we believe that the model is simple enough for us to use as prediction of prices for any listing.