Contents

1 Introduction

1.1 Background

After emerging alive from the devastating R-Town Bombings of November 2018, Spliney wanted to start a new, happier life in a different town. Him and Poly considered a number of different cities: SQL City, Javaville, and even C++ Cove but none of them seemed quite right. They wanted to go somewhere more exciting, more romantic, more real. They wanted to go to Paris: *The City of Love.*

After two years of living happily in separate apartments in Paris, Spliney and Poly decided to take the next step and move in together. Both owning their apartments, the two had a big decision to make; which apartment would they live in, and what would they do with the other one. Lucky for Spliney, Paris is not only The City of Love, it is also *The City of Airbnb.* According to Statista, Paris has over 60,000 properties listed on Airbnb 30,000 more than the second most popular Airbnb city, London. This presented an opportunity; renting out the second apartment as an Airbnb would be a great way for the young couple to generate some supplemental income.

1.2 Goals

To help our friend Spliney, we will be investigating the different characteristics of a Parisian apartment to maximize monthly revenues on Airbnb. Specifically, we will analyze which characteristics contribute to income generation, so Spliney and other individuals can better understand the most and least important areas of investment. These include, but are not limited to: property type, size and amenities that are offered by the Airbnb. We will also analyze the location of potential properties, a variable that is especially well suited to the geography of Paris. Paris is divided into 20 distinct administrative districts (arrondissements), each with its own unique identity and zip code, which will make it possible for us to directly compare between properties in different arrondissements.

Our end goal is to create an model where users can input data about their property and receive a monthly revenue prediction. Further, users (such as Spliney) can also

use the model to compare how much adding a certain amenity will improve their monthly revenue. This will enable property owners to better understand their expected revenue, and where to invest to create the greatest improvements to their monthly yield. By helping users understand which factors are most likely to improve their revenue and by how much, we can help them to achieve the highest revenue possible with the most efficient capital investment. This will help many individuals to create an optimal Airbnb listing in Paris.

2 Data Description

2.1 Data Cleaning

We started our process by finding a dataset of over 60,000 Parisian Airbnb listings, consisting of a multitude of continuous and categorical variables from average cleanliness rating to whether or not the property offered shampoo. However, one of the first things we noticed was that many of these data points did not include the size (square metres) of the property. We hypothesized that the size of a property would be a pivotal variable in predicting revenue per month, so we decided to first eliminate all data points that did not include this information. By removing the data points that did not include size, we narrowed our dataset down to a mere 951 data points. The final dataset was far more condensed than expected, however we concluded that it was essential to account for the size of the property when making our predictions.

Next, we wanted to ensure that our condensed sample was representative of the overall population of Airbnbs in Paris. In order to verify this, we conducted a series of Single Sample T Tests to compare the characteristics of our smaller sample with the initial sample of over 60,000 listings. Although we looked at a number of characteristics including p values and standard deviation, we were most interested in the means of both the initial dataset and our selective sample. Luckily, we found the datasets to be quite similar across a number of key aspects. For example, our smaller sample had a mean Average Airbnb Rating of 92.847 (out of 100) compared to 92.813 for the entire dataset. Similarly, our sample had a mean Nightly Price of 119.35, very closely resembling the entire datasets mean of 116.72. The similarities

between our sample of 951 listings and the entire Airbnb dataset reassured us that our data was representative of the entire landscape, and would be integral in structuring a predictive model.

2.2 Data Imputations

Upon taking a holistic view of our dataset, we realized that there were some important variables that were not yet accounted for, namely occupancy rate and revenue per month. We decided to create proxies using assumptions and research to estimate the values of these variables. To estimate occupancy rate, we worked with a variable that we already knew number of reviews per month. Through a series of anecdotal evidence (one of our group members hosted 78 guests on Airbnb this summer, and another works at a property rental company in Paris recently purchased by Airbnb), coupled with online research, we deduced that roughly 50 percent of guests leave reviews on Airbnb. From this knowledge, we decided to take the number of reviews per month and multiply it by two, to estimate the number of total bookings per month. We then multiplied this number by 5.2, to account for the fact that most guests rent out Airbnbs for an average of 5.2 nights in Paris, according to Inside Airbnb. We then divided that number by 30, to get the estimated percentage of days in a month that the listing was likely to be occupied. For example, a 50 percent occupancy rate represents 15 days of occupancy per month. Creating this new variable allowed us to estimate an occupancy rate for all of our data points, which will in turn help us in estimating monthly revenues.

The second proxy that we created was for revenue per month. For this imputation, we utilized the previously calculated occupancy rate. To calculate the monthly revenue at full occupancy, we multiplied the price per night times thirty. We then multiplied this value by the occupancy rate, to properly estimate average monthly revenue per property. Creating this new variable was crucial to understanding and predicting average monthly revenues with our model.

Upon examining this new variable, we decided to remove all data points that had a revenue per month that was less than a single nights price. This is because these were outliers that our proxy had likely estimated inaccurately. Removing these

points ensures that our data is representative and not skewed due to data points with overly small monthly revenues. In contrast, we also looked to see if there were any data points with monthly revenues above the price of staying 30 nights. There were not, so we did not need to remove any data points for this reason. Additionally, we removed all data points that still had N/A values, further reducing our dataset down to 691 data points.

Next, we created a new variable to account for the bias of a larger Airbnb. Our new variable, Revenue per Square Meter, accounts for the size of an Airbnb in order to make more accurate predictions. Without accounting for size, larger Airbnbs would be at an advantage. This variable ensures that our model considered the size of an Airbnb in order to make the most integral predictions possible. We created this variable by dividing revue per month by square meters.

Furthermore, many of our variables were categorical in nature, such as whether or not a listing has wifi or a television. To appropriately analyze these variables and compare them across the various listings, we created dummy variables to represent each categorical characteristic. This allowed us to conduct a similar analyses to what we did for the midterm project with movie genres efficiently isolating one group and using it as a benchmark for which to compare the rest of the data. By implementing dummy variables, we were able to take a holistic view of the dataset and analyze all important variables.

2.3 Analysis of Variables

After adjusting the data and creating dummy variables, we examined the relationships amongst variables to better understand correlations before building our model. We noticed many interesting things, for example that roughly 40 percent of listing accommodate two people, and roughly 25 percent accommodate four people (Appendix 1). Interestingly, less than 10 percent of listings accommodate three people. At first we found this the be puzzling, but then we thought about this from the point of view of an Airbnb owner. We hypothesize that most hosts make the assumption that one bed will accommodate two people, therefore resulting in the majority of listings accommodating an even number of people.

We also noticed that around 50 percent of hosts respond within a few hours or less (Appendix 2). When plotted against revenue per month, hosts that respond faster generally have higher occupancy rates, leading to higher revenues. Another interesting thing we noticed was that two zip codes made up almost 25 percent of all total listings, the other listings being very fragmented amongst other zip codes (Appendix 3). These two zip codes were 75011 and 75018. 75011 (the 11th arrondissement) is close to the city center, leading to high demand and a significant percentage of total Airbnb population. The 18th arrondissement is close to Sacr-Cur, slightly north of the city center. This is less expensive, attracting many economically inclined guests, accounting for another significant portion of total listings.

Looking at other predictors, we discovered that the number of people that a property accommodates has a strong impact on the monthly revenue, as we expected. We were also not surprised to find that the zip code had a strong effect on monthly revenues, as some areas in Paris are simply more expensive and popular, resulting in a higher revenue per month. When creating an Airbnb, it would make sense for properties that can accommodate more people and are in nicer areas to bring in higher monthly revenues.

In looking at the average monthly revenues per district (Appendix 4), two districts stood out: 75001 and 75013. The 1st arrondissement, the district closest to the city center, has the highest median monthly revenues. This is what we expected, as this district is quintessentially Parisian, and one of the most iconic areas in the world. Encompassing attractions such as the Louvre, and the Tuileries Palace, it truly is the heart of Paris. The 1st arrondissement is also the smallest district (in both area and population) resulting in a high demand and low supply, driving up monthly revenues for Airbnb properties. The 13th arrondissement is the district with the lowest average monthly revenues, which is understandable given its distance from the city centre, and relative lack of economic prosperity.

Another variable that had the largest positive effect on revenues was surprising; Instant Book. We did not expect for this factor to have such a large impact on monthly revenues. This factor is something that can be easily implemented by most

hosts, and is something that we can suggest to Spliney and other individuals on our app. We were surprised to find that the most impactful amenity was whether or not a place had an iron. This was more positively correlated with higher revenues than even having a TV, something that we found surprising. Roughly 67 percent of listings do have an iron, and this amenity is something that we would suggest to Spliney and other future hosts to invest in.

3 Model Selection

3.1 Methodology

Before building our model, we decided to use logical reasoning to make sure that all of the variables included made sense. We had to remove any variables from our model that individuals would not be able to input when considering a property, in order for our model to be realistic. For example, we removed response rate as individuals would not know this information prior starting their Airbnb. We kept variables such as zip code and number of people a property accommodates, as users would be able to input this information into our app, thus we can use it to predict their monthly revenue.

We decided to build our model using the tree-based methods of random forests and decision trees. We decided to use random forests in order to account for the possible bias that multicollinearity can create. This was important as we suspected some factors, such as square meters and number of bedrooms to be at risk of being collinear. Random forests are also one of the most powerful predictive techniques that we have learned in this course, allowing us to better understand which predictive variables are most important in making our estimations of revenue per month. Random forests often outperform simple decisions trees as well as linear and polynomial models, and are not susceptible to overfitting to the training set. Additionally, we examined decision trees to gain a visual understanding of a possible tree in the forest, and see the improvement that a random forest provides over a simple decision tree.

In addition to utilizing random forests, we employed clustering to identify a series of distinct subgroups within our dataset. By utilizing the K-Means clustering

technique, we discovered three distinct clusters when plotting the occupancy rate of all our properties (Appendix 5). Based on our knowledge and the clusters, we identified three different Airbnb host commitment levels: occasional, regular, and full time. Using K-Means clustering, we determined that hosts with an occupancy rate below 20 percent were regular hosts, perhaps only wanting to rent out their property on the weekends. Hosts with occupancy rates between 20 percent and 60 percent were occasional hosts, looking to rent out properties when it is convenient. Finally, hosts with occupancy rates above 60 percent could be classified as full-time, looking to rent out their Airbnbs as much as possible.

By clustering our data into these three segments, we were able to improve the MSE of our model by 50 percent. By capturing a hosts commitment to success, our model was able to make much better predictions, as the average monthly revenue of a occasional host is much smaller than that of a full-time host. These cluster segments ended up being the most impactful predictor on revenue per month. Additionally, commitment level is something that users can input into an app, where as occupancy rate is something that cannot be provided prior to starting an Airbnb.

We then calculated the predictions based on the test set, and used MSE to measure our results. However, we found our MSE score to be very inflated due to the large range of our response variable, revenue per month. Unlike our midterm project where the IMDB ratings ranged from 0 to 10, our response variable for this model ranges from 28.80 to 12,533.92. We felt that our MSE was overly inflated due to this large range, and therefore decided to measure revenue per month by thousands, for example an revenue of 1200 became 1.2. This provided us with a much smaller range; .0288 to 12.5339, allowing use to derive a more representative MSE score. The MSE of our final model is .4157, or an error 415.7 on average.

3.2 Model Construction

We started by building a random forest using all available inputs; 76 different variables. This gave us an idea of what the most powerful predictors were, and we selected the 16 best predictors to include in our model (Appendix 6). When choosing the predictors, we sought to balance our models predictive power, with

greater simplicity and ease of use in our app. Regarding predictive power, we looked for variables that lead to the largest percentage decrease in MSE, reflective of their ability to improve our models accuracy. The top three most powerful predictors of revenue per month are host commitment level (derived from clustering), the property size in square metres, and the number of guests accommodated. This is not incredibly surprising, as it explains the number of nights an Airbnb will be available, and the amount of space it has to accommodate guests. Looking more holistically at the predictors, we thought about our model from the user's perspective, and eliminated variables that would be difficult or irrelevant for users to input. For example, we removed a dummy variable on whether or not the property has a loft, as it would not be applicable to most users. Treating our model construction as an art not a science, we are confident that we selected the best combination of predictors for our model to have high predictive power and ease of use. After selecting our predictors, we split the data into a training and a test set.

We used 75 percent of the observations in the training set, and the remaining 25 percent were included in the test set. Before creating our final random forest, we trained a simple decision tree using our training set, in order to create a baseline for reference. We created the simple decision tree by first creating a tree with a very low CP threshold, essentially overfitting by creating splits that only lead to minor predictive improvements. We then found the CP threshold that creates the optimal decision tree with the lowest MSE, to use for our analysis going forward. Through this method, we were be able to compare our random forest results with that of a simple decision tree, providing a valuable benchmark. Additionally, we gained a visual understanding and were able to see the types of decisions happening at each split in the tree (Appendix 7).

We then trained the random forest using the training set, and inputted the test set to create predictions that were generated by our random forest. We used this method so that we can test our models predictions with data that the model has not seen before. This accounts for training bias and overfitting. Within our random forest we create 500 trees, and at each split considered a random subset of five out of the 16 variables. This ensures that our trees are not correlated, while still incorporating all variables. Finally, we compared the predictions from our random

forest to those of the simple decision tree, and were pleased to see that our random forest produced significantly better results.

4 Results

4.1 Recommendations for Spliney

Now that weve built our model (Appendix 8), we can help our friend Spliney and other individuals to invest in the best Airbnb properties possible, and understand important factors to improve their properties. Deriving data from our distilled sample of 691 properties, we were able to explain 72.57 percent of the variance in monthly revenues, arriving at a mean squared error of residuals of 0.041. We will now use this model to predict monthly revenues for our friends Spliney and Poly.

In order to help our friends, we need to compare the two properties they are considering listing on Airbnb one that belongs to Spliney, and one that is owned by his girlfriend Poly. Spliney and Poly love both of the apartments are are indifferent to which one they live in, caring only about maximizing their monthly revenues from the Airbnb property. We will help them by predicting the monthly revenues for each of their apartments, so they can decide which apartment they should live in and which they should rent out. The features of their apartments are as follows:

Feature	Spliney	Poly
Property Type	Entire Appartment	Entire Appartment
Zip Code (Arrondissement)	75001 (1st)	75015 (15th)
Family Friendly?	No	Yes
Size (Square Meters)	32	53
Commitment Level	Full-Time	Full-Time
Guests Accomodated	2	4
Number of Bedrooms	1	2
Number of Bathrooms	1	1.5
Number of Beds	1	2
TV?	Yes	No
Shampoo?	No	Yes
Air Conditioning?	Yes	No
Desk?	Yes	No
Iron?	No	Yes
Hair Dryer?	No	Yes
Kitchen?	Yes	Yes
Predicted Income (Euros)	**2334.90**	**2604.60**

Table 1: A table comparing the characteristics of Spliney and Poly's apartments. Spliney's apartment is in a more prestigious arrondissement, the 1st, but Poly's is much larger and can accommodate twice as many guests, making it a more lucrative property to list on Airbnb.

Upon inputting this data from Spliney, our model predicts a 354.40 discrepancy between monthly revenues of the two listings. The couple can make 15 percent more income per month by renting out Polys apartment, for a total of 2604.60. Despite Splineys property being in a more desirable neighbourhood, the first arrondissement, Polys apartment can generate more revenue on Airbnb because of its larger size and capacity to accommodate twice as many guests.

In order to maximize the revenue our friends are able to generate, we recommend a few changes to Polys apartment before listing it on Airbnb. By adding a TV she can increase monthly revenues by 234, while purchasing a desk and installing an air conditioner will generate an addition 92 and 127 per month respectively. Although adding these amenities will have an initial cost, the additions will increase monthly revenues by a total of 453, increasing Spliney and Polys earnings by roughly 25 percent.

4.2 Model Applications

Looking beyond the scope of our friend Splineys predicament, our model can provide incredibly valuable information to potential Airbnb hosts. According to Business Insider, Airbnb has reached over 5 million property listings, more than the top five hotel brands combined! With the hospitality industry undergoing a seismic shift from traditional hotels to home and apartment rentals, it is critical to know which rental properties will perform best. For example, a university student could utilize our model to find the expected revenue from renting out their apartment when they return home for the summer. To maximize that figure, they could look into the most cost effective measures to increase monthly revenue. For example, adding a TV is far more affordable than installing an air condition unit, and increases Airbnb revenues by a larger amount. Similarly, families could employ our model when renting out their homes during family vacations.

Our model could also be used by property management companies, who own and operate a large number of properties with the sole intent of generating revenue. One example is Sonder, a hospitality startup founded by McGill students. Sonder aims to rent out apartments (both on their own website and on Airbnb) with hotel-level amenities. Companies could use our model when deciding which properties to acquire, and better understand how to properly outfit them for renting in order to maximize their success as a business.

4.3 Limitations and Next Steps

Despite its ability to predict monthly Airbnb revenues in Paris, our model did have some limitations that we would want to address in a future model. First off, when building our model, we had to remove most of the data points due to lack of information on the square meterage of each apartment. In the future, we would want to build a model using a more complete dataset, in order to utilize more data points and have a larger sample of evidence backing our predictions. Further, it would have been ideal to have a larger sample of data from other types of accommodation, primarily shared apartments, to compare against entire properties up for rent.

Secondly, our model addresses monthly revenues, but is unable to generate project profits. With more information and time, we would have wanted to investigate the capital requirements of each property to provide property owners with a better understanding of profit rather than revenues. Some costs we would be interested in investigating include property taxes, utilities, cleaning fees, and maintenance costs, that could vary depending on the size, location, and characteristics of each property. With all of these costs in the fold, wed be able to give model users a better indication of how much profit theyd be able to generate from renting out a given property. It would be fascinating to see if some properties have costs that eclipse monthly revenues, and should be avoided altogether.

Third of all, our model was limited in its ability to capture creative pricing strategies that hosts are able to employ. For example, some hosts will charge a lower nightly rate, with a higher cleaning or transaction fee. We would be interested in knowing if this is the best path to maximum revenues, or if it less effective than simply charging a higher nightly rate. Many Airbnb hosts also have a minimum stay duration in order to ensure a certain threshold of revenue from each guest. In a future analysis, wed like to explore if this is an effective tool for obtaining high occupancy rates, or if it alienates too many potential customers?

Finally, in future statistical analysis, wed want to corporate an element of seasonal and cyclical demand into our model. For example, some districts, likely those with lots of parks and green-space, will be especially popular during the Spring and Summer months, but less so in the Autumn and Winter. Knowing the seasonal distribution of both occupancy and pricing for each apartment would help us better understand the market environment, and make more informed predictions on monthly revenues at any given point in time. Looking to cyclicality, we would like to adjust our revenue projections based upon economic cyclicality, particularly economic growth and the health of the tourism industry in Paris.

5 Appendix

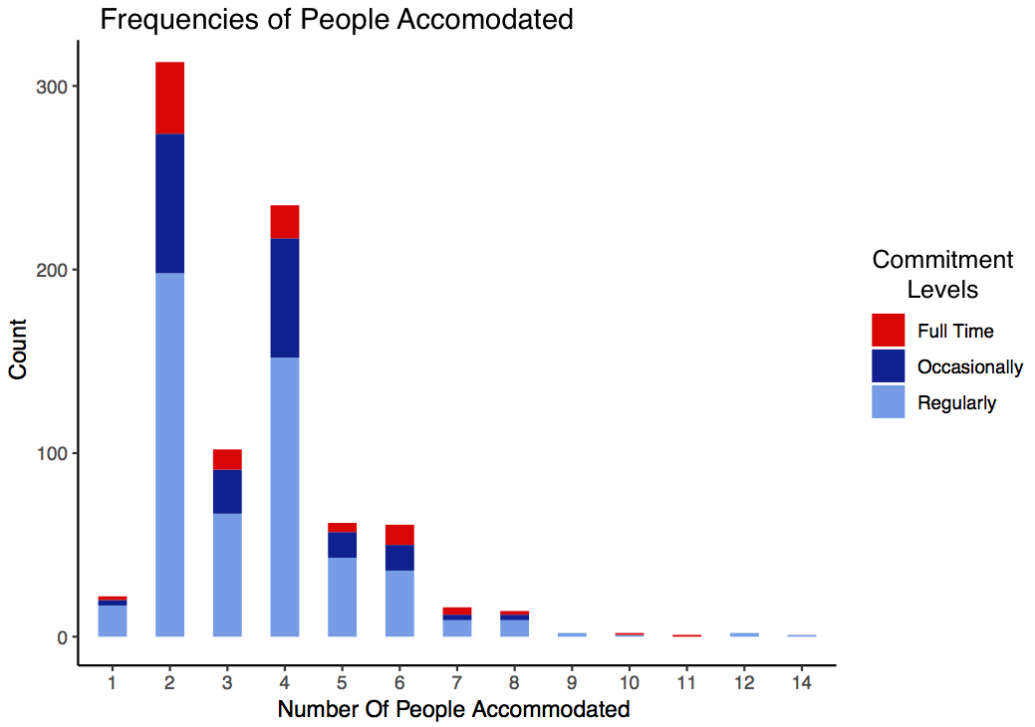

Figure 1: A histogram showing the number of Airbnbs in Paris that accommodate each number of guests. We see that the most frequent are apartments designed for two and four, followed by Airbnbs that can accommodate three, five or six guests. Very few can only accommodate one visitor. Further sub-dividing by commitment level (is the Airbnb listed full-time, regularly, or just occasionally), we can see that a disproportionately large number of two person properties are listed listed full time or occasionally. **Note:** Throughout this report we will refer to full time hosts as the most committed, occasional hosts as the second most committed, and regular hosts as the least committed.

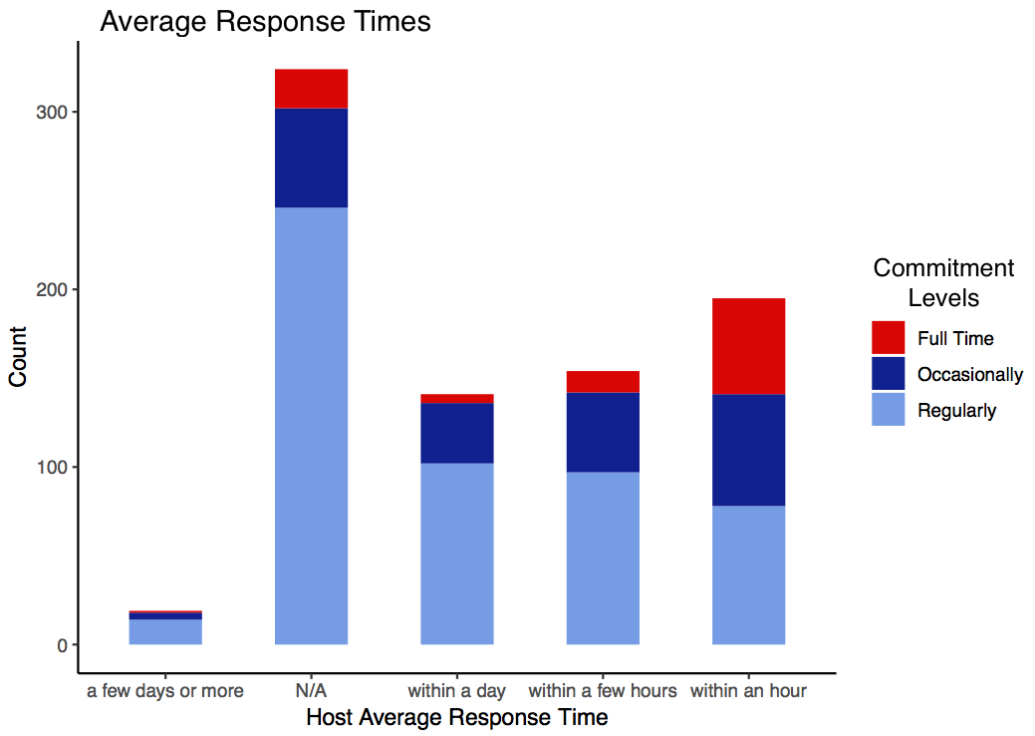

Figure 2: A histogram showing the distribution of host response times to messages on Airbnb. Interestingly, among hosts who choose to respond, almost half respond within an hour, and the vast majority respond within a couple of hours. Among the hosts who respond within an hour, around a quarter are full time hosts, who respond faster and show an increased level of commitment to the listing.

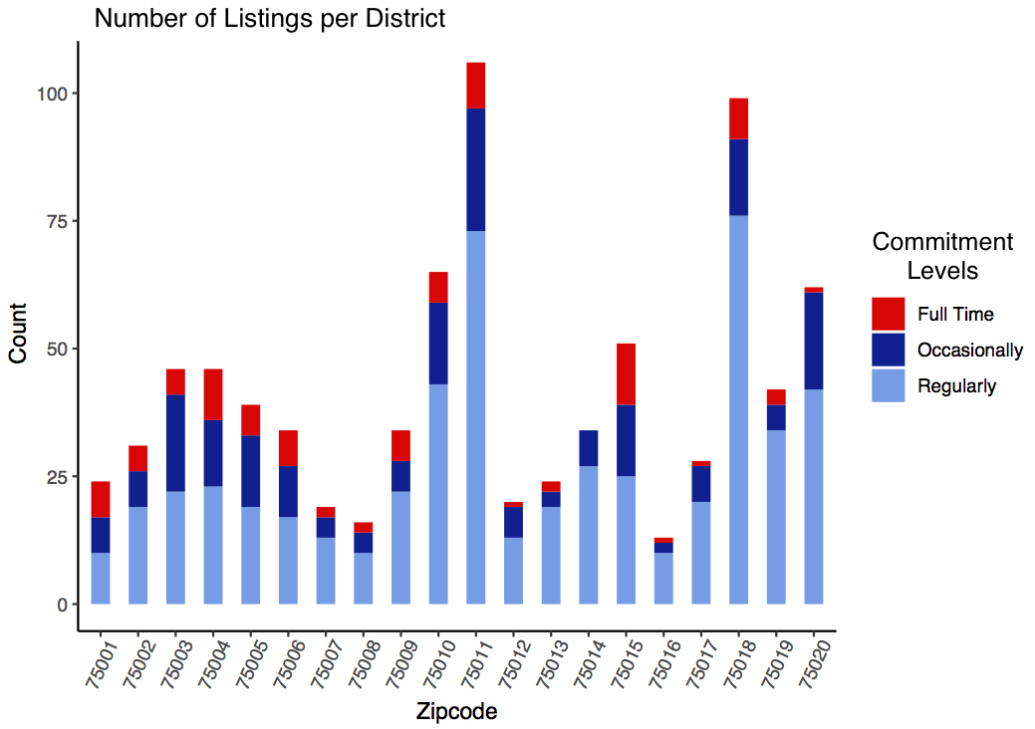

Figure 3: A histogram showing the number of Airbnb listings in each district of Paris. We can see that the 11th and the 18th, two of the more populous districts, have the most listings. Converseley, the most central districts (1st, 2nd, 7th, 8th) are far smaller and therefore have fewer listings.

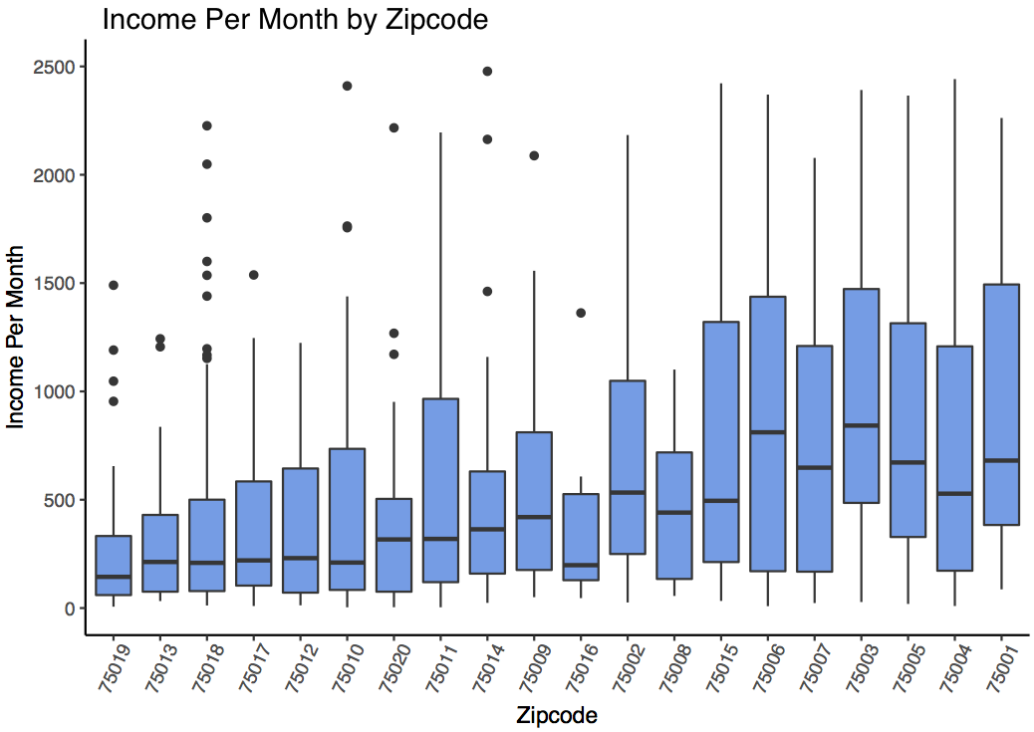

Figure 4: A box plot visualizing the distribution on monthly listing income within each zipcode (district), and between the different districts.

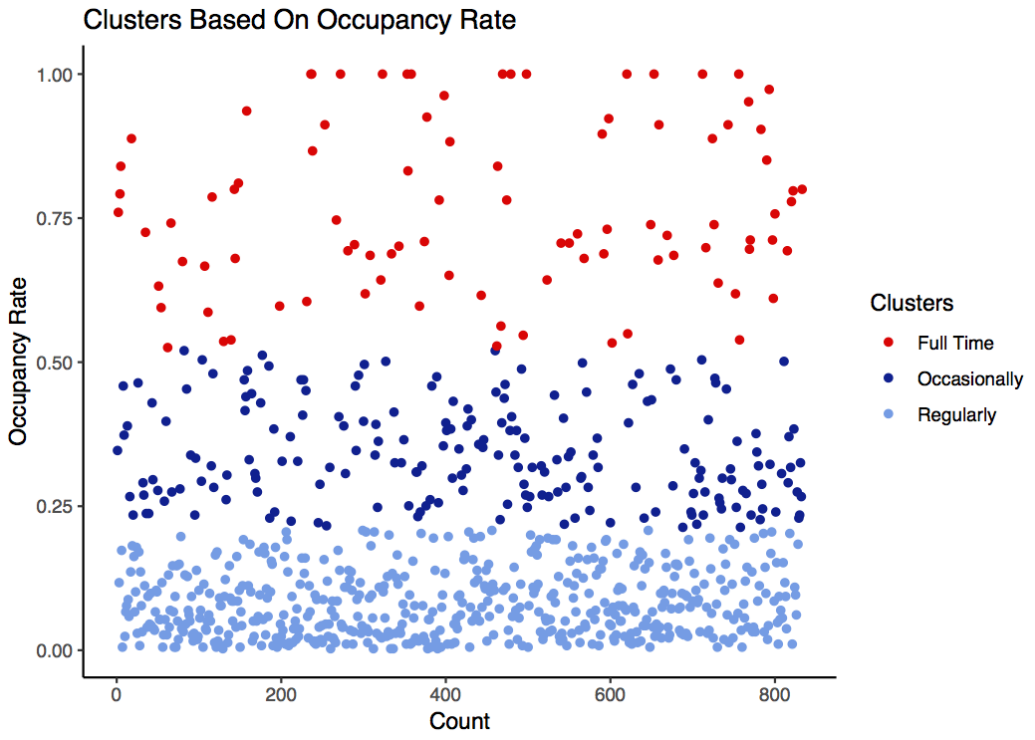

Figure 5: A scatter demonstrating three distinct clusters of Airbnb listings by the host commitment level. The hosts with the highest occupancy rate are shown in red, and designated as full-time hosts, who have an occupancy rate of over 60 percent. The occasional hosts shown in dark blue have an occupancy rate of 20 to 60 percent, while the light blue regular hosts have an occupancy rate of under 20 percent. We see that the full time hosts are the least common, followed by occasional, and then regular.

Variable Name	**% Increase in MSE**
Host Commitment (from Clustering)	51.11
Square Metres	14.36
Guests Accommodated	10.47
Number of Bedrooms	8.891
Number of Bathrooms	7.412
Number of Beds	5.750
Iron?	4.973
TV?	4.102
Desk?	2.793
Family Friendly?	2.635
Air Conditioner?	2.573
Hair Dryer?	2.160
Zip Code	1.216
Kitchen?	0.448
Property Type	0.146
Shampoo?	0.021

Table 2: A table displaying the 16 predictors we decided to include in our model, ordered by their impact on MSE. Variables with a higher '% Increase in MSE' increase the standard error by a greater amount when removed from the model. For example, when Host Commitment, the most important predictor, is removed from the model, the MSE increases by an astounding 51.11 percent. This is in stark contrast to the Shampoo variable, with a far smaller effect on MSE.

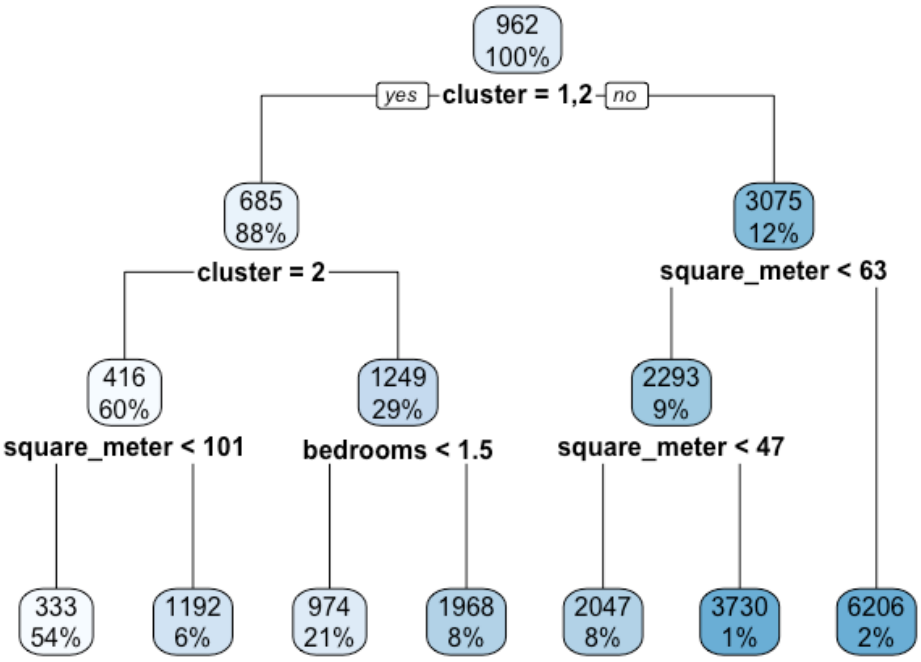

Figure 6: An optimal decision tree demonstrating the variables considered at each plit, starting with all of the properties, and eliminating

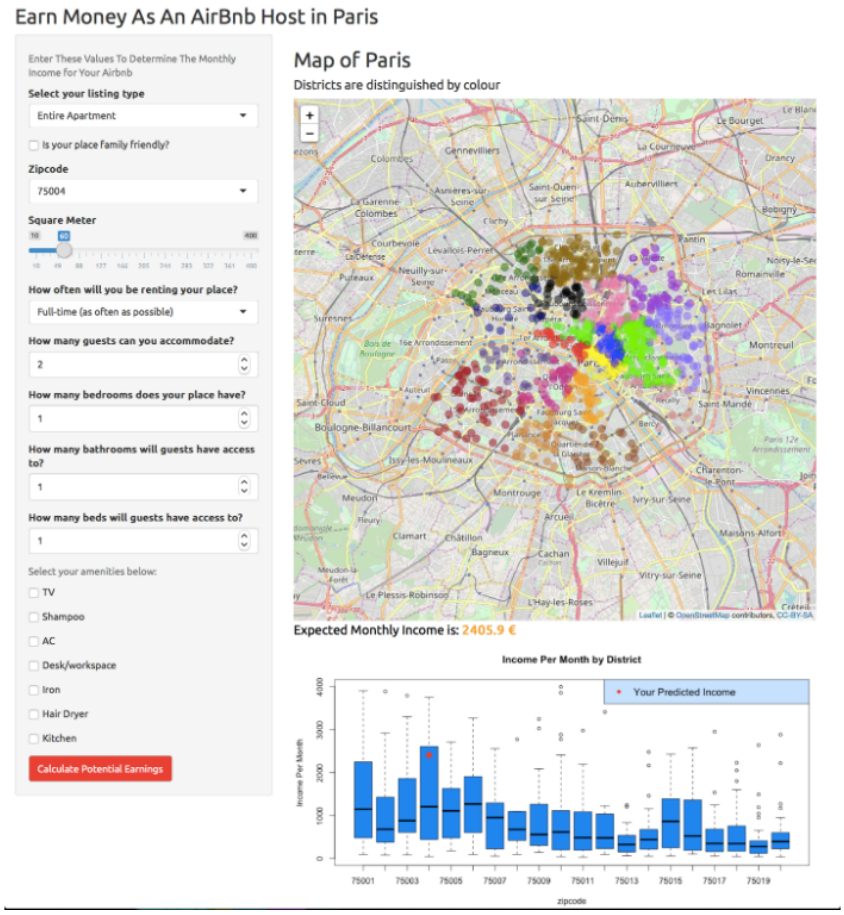

Figure 7: A screenshot of an application we created to determine the monthly revenue of a Parisian Airbnb listing given its location and characteristics. We built this interactive platform using R Shiny in order to allow users to enter their own inputs more efficiently. The map shows a distribution of Airbnbs in the different districts of Paris, while the boxplot on the bottom places your listing (red dot) relative to all others in Paris. The instrutions for running the application are as follows: 1) Download files: AirbnbForest.R, server.R, ui.R, Making-Graphs.R, FinalCleanedAirbnb.csv 2) Ensure all files should be in the same location 3) Open and run AirbnbForest.R (complete file - ensure FinalCleanedAirbnb.csv is loaded) 4) Open and run Making-Graphs.R up to line 6 5) Open and run server.R and ui.R (complete files) 6) Run app from server.R 7) Follow instruction on app to get your prediction!

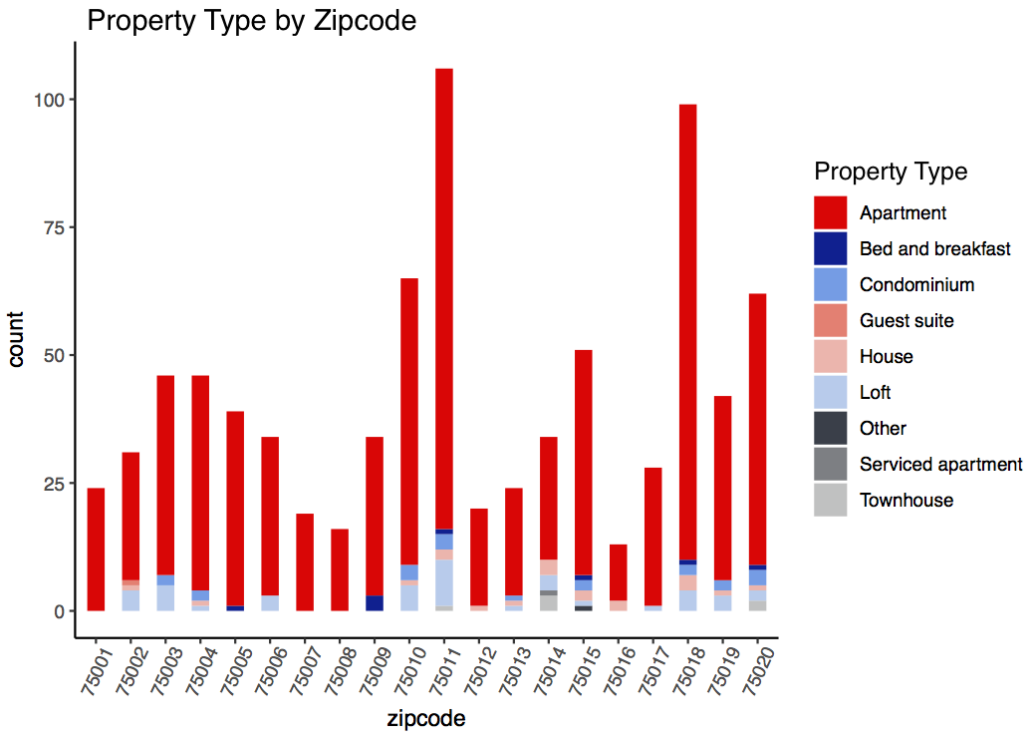

Figure 8: A histogram showing the different propert types amongst the different zip codes of Paris. Across all the districts, the vast majority of listed properties are apartments, with very few amounts all of other types.

Figure 9: A word cloud showing the frequency of various words that appear in the Airbnb descriptions from our dataset. The words Paris and apartment are the most prevalent.