****

**Post-Pandemic Airbnb Markets: New Orleans, LA**

**Sean Oberer, Matt Campbell, Deevanshu Kishor Khatri, Krishna Deepika Chavali, Ambreen Abbas**

**Introduction**

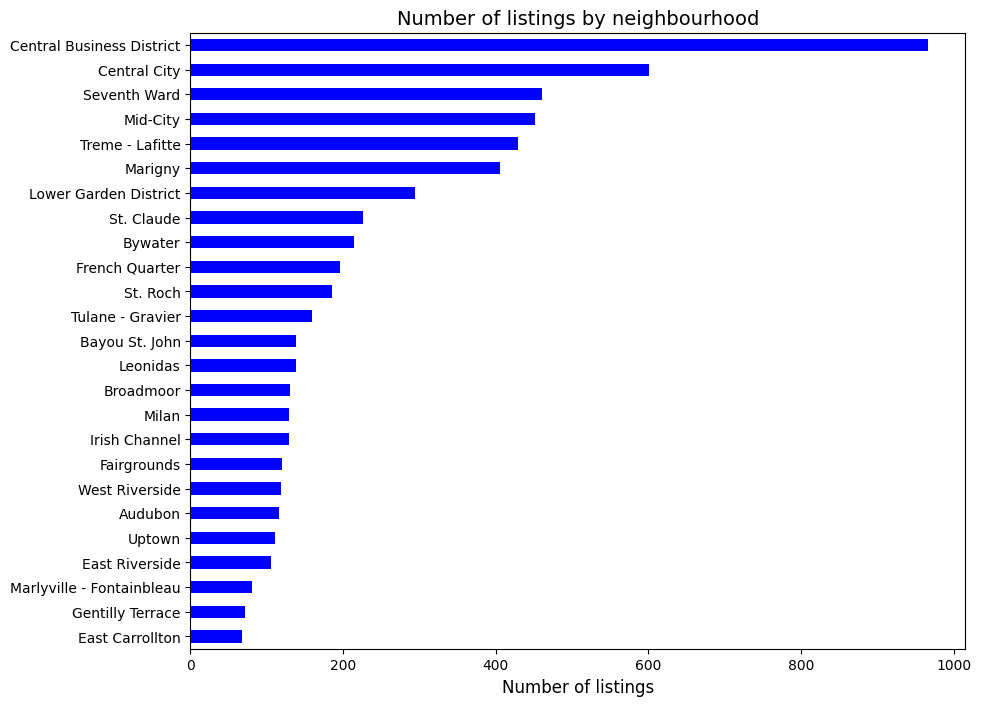
This paper will discuss the post-pandemic Airbnb market for New Orleans, LA. We will start with performing exploratory data analysis to understand the new orleans data then we will take a deep dive into the market demand, market supply, what makes an Airbnb ‘successful’ from a quantitative perspective in regards to a particular listing's characteristics, and determine what the most influential variables are when determining what the price of a particular Airbnb will be. We will also be using text mining to obtain a better understanding of the terms used in describing a listing, as well as the terms most commonly used by reviewers.

The data that was used for this project can be found here: http://insideairbnb.com/get-the-data/

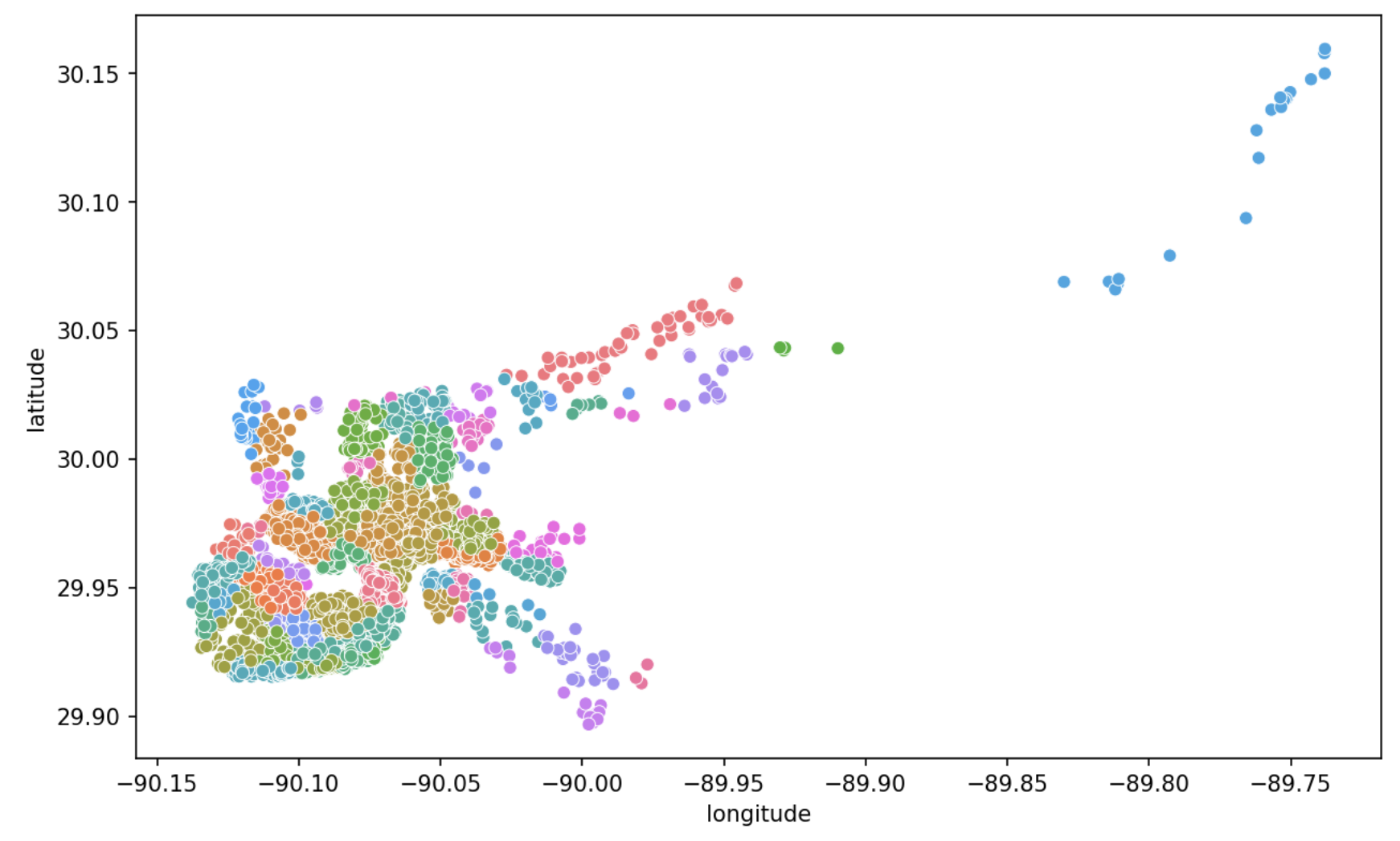
Our GitHub repository with all of our code for this project can be found here: <https://github.com/soberer/6211-Project>

**Exploratory Data Analysis**

1. **Density of Airbnb by neighborhood**

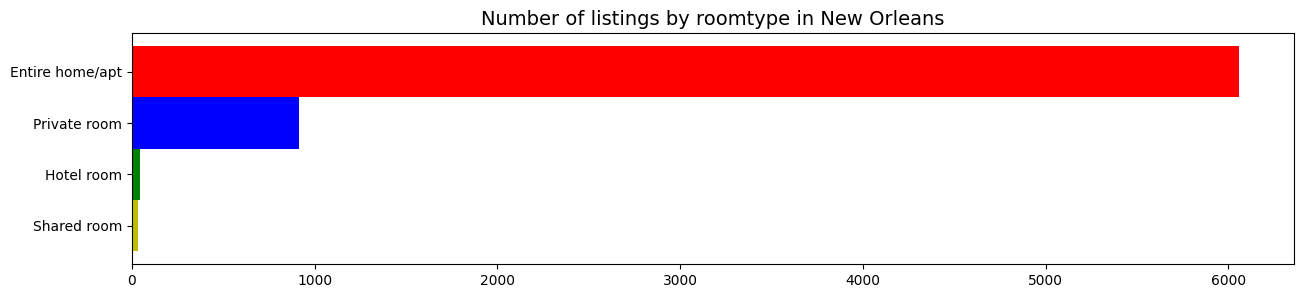
****

Neighborhoods which are nearby to tourist places have a high number of Airbnb’s which is expected. The Central Business District and central city has a significantly high number of Air BNBs as its proximity to tourist places like Jackson Square and Downtown. Seventh ward, mid city and Treme-Lafitte are closeby to jackson square and new orleans city park.



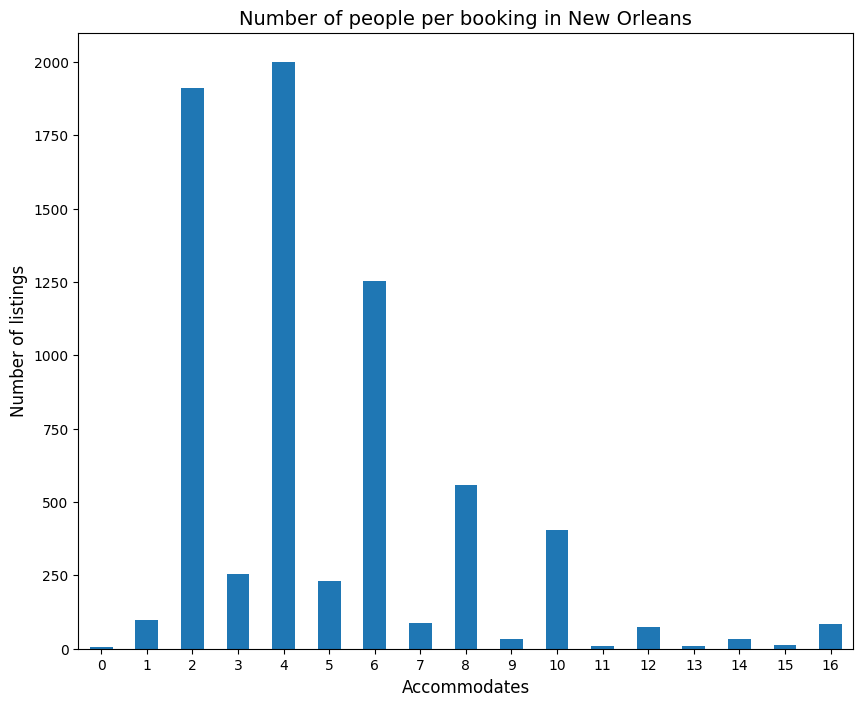
1. **Most preferred type of Airbnb**

Majority of the Airbnb’s in New Orleans are entire houses or apartments which indicates it is the most preferred type of Airbnb.



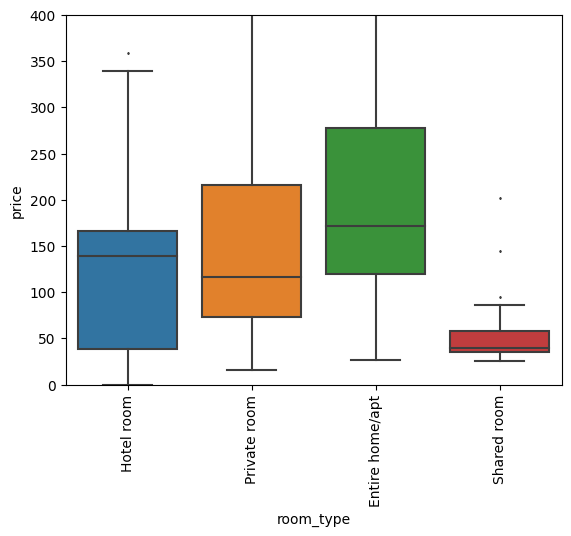
1. **Number of tourists per group**

Majority of the tourists visiting New Orleans come in a group of 2,4, or 6 indicating that New Orleans is a preferable destination for couples.

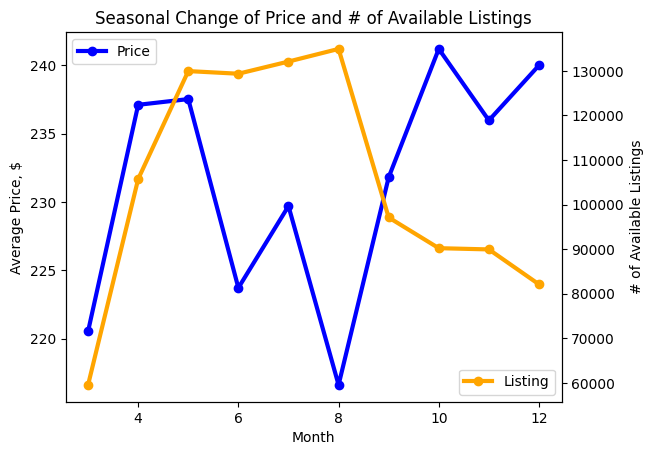


1. **Variation in Price**

Entire home or apartment has the highest price per listing followed by Private Room, Hotel room and shared room.



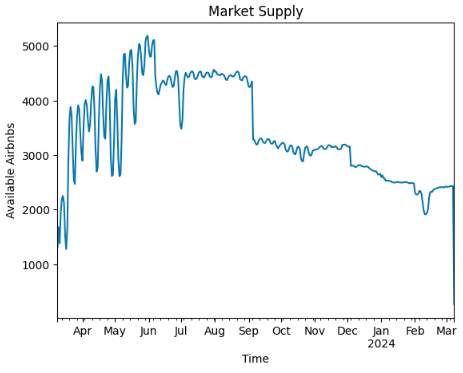
Prices are high during peak holiday season in New Orleans - Mardi Grass Mania during Feb - May and Winter Break during October - December.



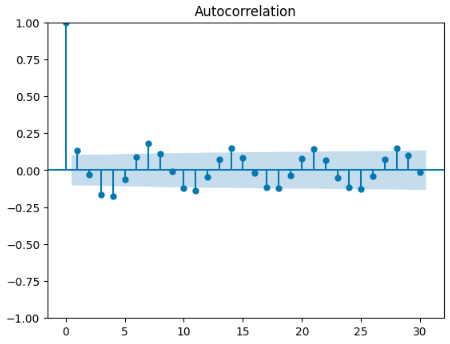
**Market Supply**

The second section that will be covered is the market supply. In this paper, market supply will be defined as follows: the total number of available Airbnb’s over a period of time. For this analysis, the ‘calendar’ dataset was used which contained detailed calendar data for a one year period. This dataset in particular contained the following variables: listing\_id, date, available, price, adjusted\_price, minimum\_nights, and maximum\_nights. The variables that were kept to be analyzed for market supply in particular were ‘date’ and ‘available’. Since the ‘available’ variable was a simple binary variable that only indicated whether a particular Airbnb was available that day, we modified the existing variable by filtering the data and only keeping the Airbnb’s that were available in the dataset, then grouped them together based on the data and calculated the total number of available Airbnb’s for each individual day.

The plot showing our market supply in our data is shown below:



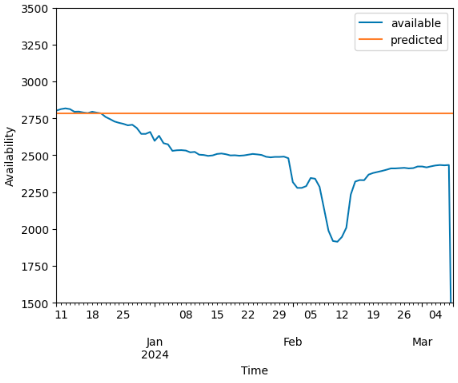
Based on our exploratory data analysis above, we can see that Airbnb’s in New Orleans has the lowest supply from December until March since the number of available Airbnb’s is the lowest in those particular months. We also see that there appears to be a quadratic trend in the data where the number of available Airbnb’s appears to increase and then decrease. Additionally, New Orleans Airbnb market appears to have the highest supply in the spring/summer from April until September since we see a high amount of available Airbnb’s. Based on the results in the figure above, there may be some seasonality based on the observations above in addition to some peak high days and peak days. After these observations, we decided to create an autocorrelation figure to see if there is any statistically significant seasonality in our data. Below you can see the autocorrelation plot created:



Based on the results above, we can see that there does not appear to be any significant indication that there is seasonality present in our data. However, we do see that there is an interesting pattern with negative and positive correlation on the same days of each week. Because of this, we will add seasonality indicators to our data and create models that include seasonality to see if seasonality does make an impact for forecasting.

As a result of these two observations, we create two more columns and add them to our dataframe to be used for time series forecasting. The first variable we introduce is ‘Day’. The ‘Day’ variable is simply an indicator as to what day of the month it is since the data we have is observed on a daily basis. For example, if the data is 03/30/2023, then Day = 30. The next variable we will introduce is ‘trend’ since we see the quadratic trend in the data above. The ‘trend’ variable will simply be an increasing constant from 1 until 366. For example, day 1’s trend value will be equal to 1, and day 2’s trend value will be equal to 2.

When creating our forecasting models, we created the following models and measured each of their performance to find which one was the best performing one to use for forecasting: model with linear trend, model with quadratic trend, a model with only seasonality, a model with seasonality and trend/quadratic trend, a moving average model, a simple exponential smoothing model, and various ARIMA models. Additionally, we used 9 (about 270 days) months worth of data for training and 3 months (90 days) of data for testing. The main evaluation metric that was used to evaluate our market supply time series models was mean absolute percentage error (MAPE). The simple exponential smoothing model performed the best with a MAPE of 22%. We can see the results of our forecasting model below:



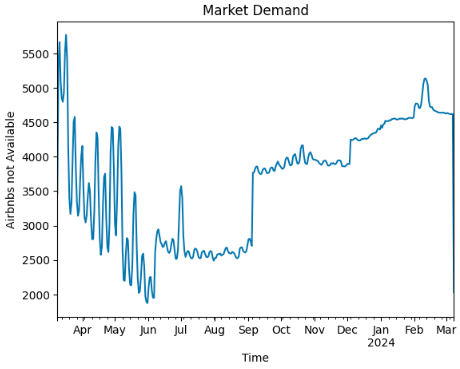
Based on our graph above, we see that our model does a decent job of forecasting the market supply but it appears to overestimate the market supply quite a bit.

Although our model was able to produce some reasonable results for forecasting, this model could definitely be improved if there were more data available with a longer time period being measured. Based on the exploratory data analysis we did earlier in this section, there appeared to be a quadratic trend and seasonality with high peak and low peak time periods. Since our dataset only contained a years worth of data, it was difficult for our models with trend and seasonality to truly capture the whole relationship between market supply and trend/seasonality.

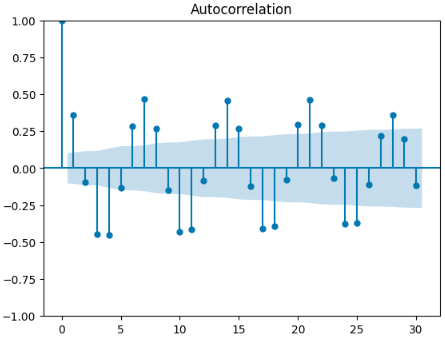
**Market Demand**

This section will cover our time series analysis for market demand in New Orleans, LA. In this paper, market demand will be defined as follows: the total number of Airbnb’s that are not available over a period of time. The variables and dataset used for this analysis is very similar to the market supply analysis and will use the ‘calendar’ dataset. For this analysis, the ‘calendar’ dataset was used that contained the following variables: listing\_id, date, available, price, adjusted\_price, minimum\_nights, and maximum\_nights. The variables that were kept to be analyzed for market demand were ‘date’ and ‘available’. We will transform the ‘available’ variable by filtering our data to only include Airbnb’s that are not available for each particular day, and then calculate the sum of Airbnb’s that are not available for every individual day. First, we will start by performing exploratory data analysis for market demand in New Orleans, LA.

Below you can see our exploratory data analysis for New Orleans market demand:



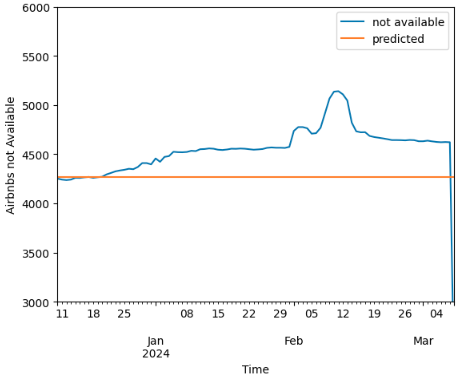
Based on the figure above, we see that there is a quadratic trend where the market demand decreases and then increase. Additionally, we see that there could also be some seasonality with some days/months having peak high demand and peak low demand. The demand for Airbnb’s in New Orleans appears to be low from April to September with the exception of some peak days and consistently high every day from December to March. Since we see some potential seasonality here in our model, we will create an autocorrelation chart which you can see below.



Based on our autocorrelation chart above, there seems to be an indication that there is seasonality present in our data since we see relatively high negative and positive correlations for the same days over a period of time. We will add seasonality indicators to our dataset and create forecasting models with seasonality to see if seasonality has an impact on our forecasting.

As a result of these two observations, we create two more columns and add them to our dataframe to be used for time series forecasting just like we did in the market supply section.. The first variable we introduce is ‘Day’. The ‘Day’ variable is simply an indicator as to what day of the month it is since the data we have is observed on a daily basis. For example, if the data is 01/20/2023, then Day = 20. The next variable we will introduce is ‘trend’ since we see the quadratic trend in the data above. The ‘trend’ variable will simply be an increasing constant from 1 until 366. For example, day 1’s trend value will be equal to 1, and day 2’s trend value will be equal to 2.

When creating our forecasting models, we created the following models and measured each of their performance to find which one was the best performing one to use for forecasting: model with linear trend, model with quadratic trend, a model with only seasonality, a model with seasonality and trend/quadratic trend, a moving average model, a simple exponential smoothing model, and various ARIMA models. Additionally, we used 9 (about 270 days) months worth of data for training and 3 months (90 days) of data for testing. The main evaluation metric that was used to evaluate our market demand time series models was mean absolute percentage error (MAPE). The simple exponential smoothing model performed the best with a MAPE of 7.5%. We can see the results of our forecasting model below:



Based on our graph above, we see that our model does a decent job of forecasting the market supply but it appears to overestimate the market supply quite a bit.

Our market demand time series model was able to produce very good results for forecasting but this model could definitely be improved if there were more data available over a longer period of time. Based on the exploratory data analysis we did earlier in this section, there appeared to be a quadratic trend and seasonality with high peak and low peak time periods. Since our dataset only contained a year's worth of data, it was difficult for our models with trend and seasonality to truly capture the whole relationship between market demand and trend/seasonality.

**What Makes an Airbnb Listing ‘Successful’**

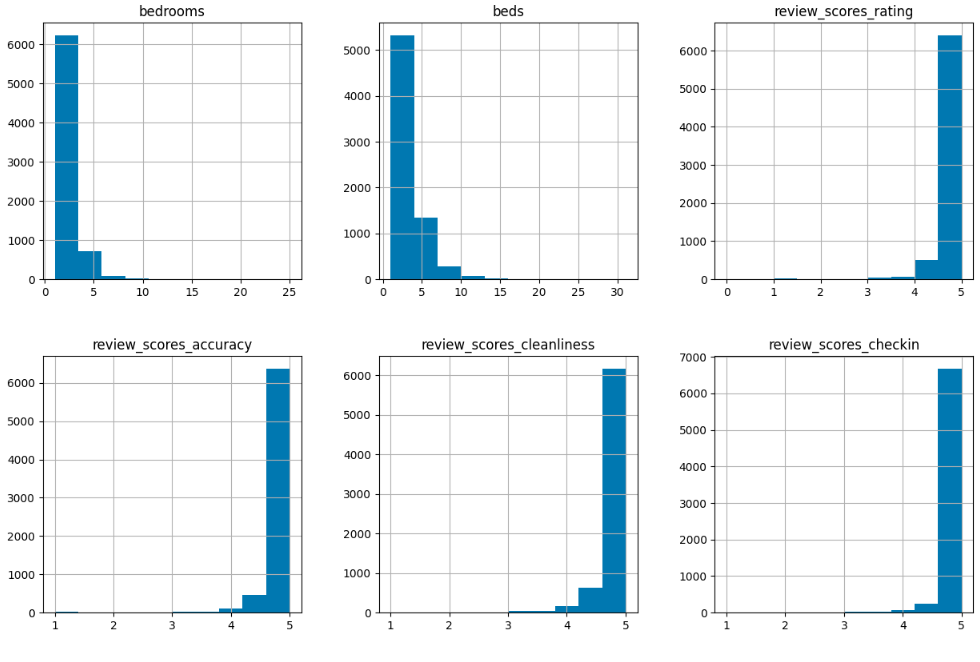
This section will cover what Airbnb’s particular individual characteristics lead to more success as a rental property. In this section, we define a ‘successful’ Airbnb as follows: an Airbnb which was booked for more than 180 days in the last 365 days. The dataset that will be used for this analysis is ‘listings.csv’. Since the problem we are trying to address was a ‘yes or no’ type of question, we created various classification models for this particular problem. The target variable is a variable named ‘successful’ which was created as a binary variable. The binary variable was constructed in a way that follows our definition of what a successful Airbnb is above - if a Airbnb was booked for more than 180 days in the last 365 days, successful = 1. If that particular Airbnb was not booked for more than 180 days in the last 365 days then successful = 0.

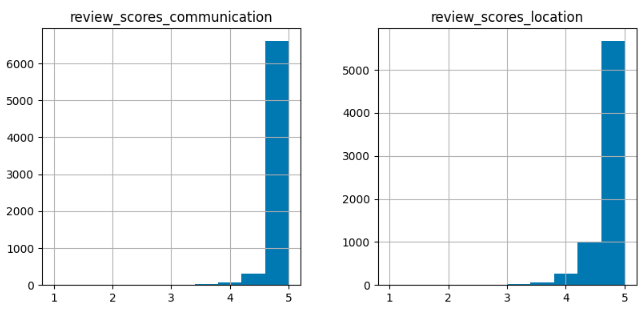
The characteristics/variables that we looked at when analyzing individual Airbnb’s were: the number of people the property can accommodate (accommodates), number of beds (beds), number of bedrooms (bedrooms), price (price), minimum nights required to stay (minimum\_nights), maximum nights that can be rented at a time (maximum nights), number of reviews in the last 365 days (availability\_365), review scores (review\_scores\_cleanliness, review\_scores\_location, review\_scores\_communication, review\_scores\_rating, review\_scores\_accuracy, and review\_scores\_checkin), property type (property\_type, such as home, townhouse, suite, etc.), type of room (room\_type, such as home, hotel, etc.), and number of bathrooms (bathrooms\_text).

The only categorical variables from the variables listed above were property type (property type), room type (room\_type), and bathrooms (bathrooms\_text). Bathrooms\_text was considered a categorical variable because it contained key information such as whether a bathroom was shared or not. The categorical variables did seem to contain a lot of unique values with low frequency so their values had to be modified to reduce the dimensionality. Since the dataset had over 7000 observations, the categorical variables that had values with less than 100 occurrences were all placed into a category specified as ‘Other’. Additionally, the categorical variables here all had either no missing values or a very small percentage of missing values, so those categorical variables which contained missing values were also placed into the ‘Other’ category. All of the categorical variables were dummy coded with the first variable being dropped to minimize the risk of multicollinearity. The dummy that was dropped from the property\_type column was property\_type\_entire\_condo. Additionally, the dummy that was dropped from the room\_type column was room\_type\_entire\_home/apt. Furthermore, the dummy that was dropped from the bathrooms\_text column was bathrooms\_text\_1 bath.

The continuous variables used for this analysis are as follows: the number of people the property can accommodate (accommodates), number of beds (beds), number of bedrooms (bedrooms), price (price), minimum nights required to stay (minimum\_nights), maximum nights that can be rented at a time (maximum nights), number of reviews in the last 365 days (availability\_365), and review scores (review\_scores\_cleanliness, review\_scores\_location, review\_scores\_communication, review\_scores\_rating, review\_scores\_accuracy, and review\_scores\_checkin). None of our columns had over 50% of their values missing so we did not drop any of these columns. However - some of our continuous columns did have some missing values so they had to be imputed.

Before doing imputation, we needed to look at the distribution of all of our continuous variables that had missing values in order to decide what imputation technique to use. The continuous variables that had missing values are as follows: bedrooms, beds, review\_scores\_rating, review\_scores\_accuracy, review\_scores\_cleanliness, review\_scores\_checkin, review\_scores\_communication, and review\_scores\_location. Below you can see the distribution of all of our continuous variables that had missing values.



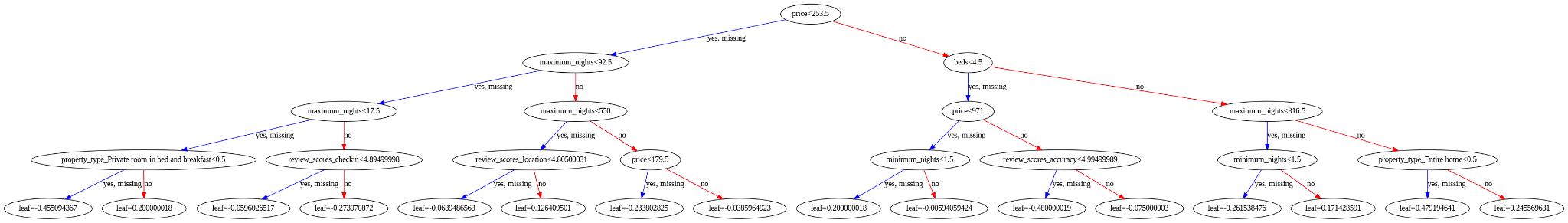


Since the distribution of all of our continuous variables with missing values were all skewed, we did median imputation for all of the variables. After we did this, our data has now been fully pre-processed and is ready to go for modeling.

Since this is a classification problem where we’re trying to predict whether an Airbnb is successful (successful variable = 1 if yes, 0 if not successful) - we created three different models: logistic regression, decision tree, and XGBClassifier. The data was partitioned into a 70/30 split so that 70% of our data was used for training the models and 30% of it was used for validation. For the decision tree and XGBClassifier models, their max depth was specified as 4 to reduce complexity of our models and prevent overfitting. The evaluation metrics used to evaluate each of the models were accuracy, precision, recall, F1 score, and area under the curve (AUC). Below, you can see our evaluation scores for all of the models we created.

| Model: | Accuracy | Precision | Recall | F1 Score | AUC |
| --- | --- | --- | --- | --- | --- |
| Logistic Regression | 60% | 59% | 37% | 46% | 61% |
| Decision Tree | 62% | 60% | 54% | 57% | 62% |
| XGBClassifier | 69% | 68% | 60% | 64% | 68% |

Based on the table above, our XGBClassifier had higher accuracy, precision, recall, F1 score, and AUC values than our logistic regression and decision tree models - making it the best performing model. As a result of this, our XGBClassifier model was chosen to be analyzed and derive useful insights/takeaways. You can see the tree plot of the XGBClassifier model below:



The figure is very difficult to see here so because of that, we propose two alternative options so that you can view if needed. The first option is to simply view the notebook (successful\_airbnbs.ipynb) located in the GitHub repository (https://github.com/soberer/6211-Project). The second option is to view the uploaded photo of the tree plot named ‘xgbc\_tree.png’ which has also been uploaded to the GitHub repository listed above.

Based on our XGBClassifier model, lower priced Airbnb’s seemed to be more successful overall. Additionally, location was crucial to an Airbnb’s success and had the highest impact on a property's success among all of the other review metrics. Furthermore, Airbnb’s that categorized as a bed and breakfast were more likely to be successful compared to other properties. Lastly, smaller properties were more likely to be successful than big properties.

**Text Mining**

An area that we wanted to further explore is the AirBnB reviews and what is said in them. This is valuable information because it allows us to see what reviewers put emphasis on. Thinking of this from the perspective of an investor, if you know what reviewers value you can use that information to reverse engineer your property to fit those values. For example, if location is frequently brought up amongst reviews, you know that it is imperative to purchase a property that is in a desirable location. You could also use this information to tailor amenities to those that are frequently mentioned in reviews. These are just two potential ideas of ways that text mining information can be used in order to improve an AirBnB listing.

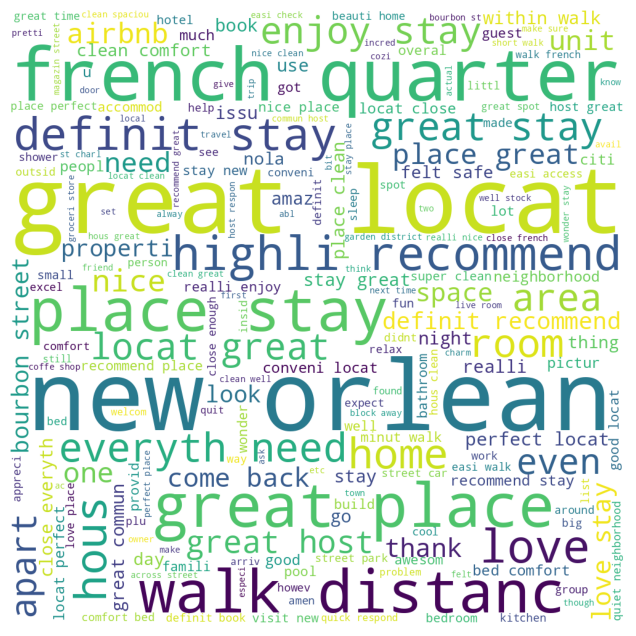
Text mining experiments were focused in three main areas: reviews from January 1st, 2023 up until the most recent review, reviews from the summer of 2022 (dates are May 1st, 2022 - September 30th, 2022) and listing descriptions from all properties with a description. The data available dates back to 2010, but we felt that was too long ago to include in this research. We instead focused on more recent listings to give a better idea of what is currently happening in the market. The data from 2023 is the most recent data so we wanted to analyze that. We also chose to further analyze the summer months from 2022 due to the trends noticed when doing time series analysis.

**Reviews Text Mining**

One of the first tasks we wanted to accomplish was to visualize the most frequent words found throughout reviews. A WordCloud was created to accomplish that. In the figure below you can see the WordCloud for the 2023 reviews data:

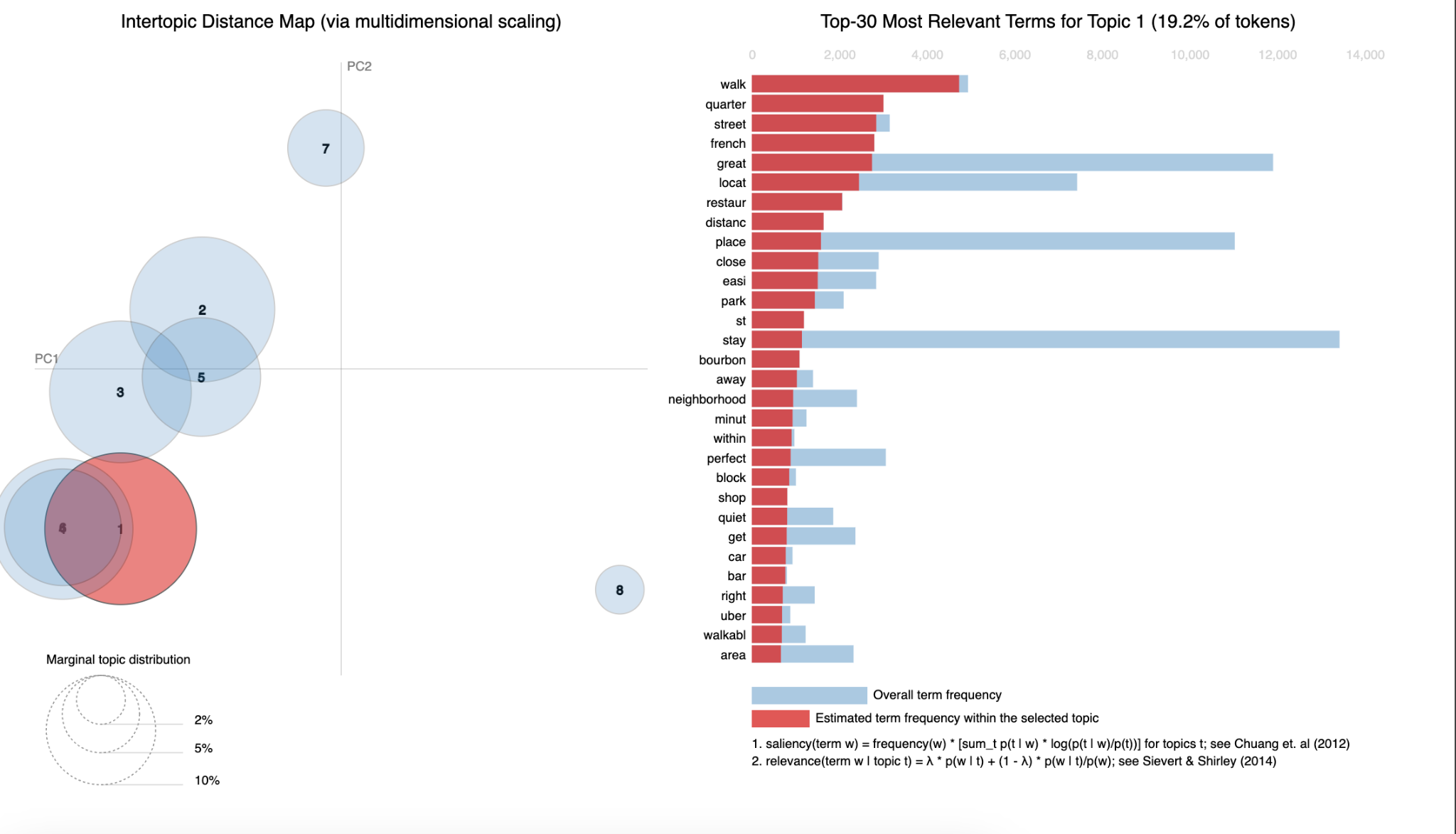


A few things that stick out is the amount of words that relate to location. Terms like french quarter, walk distanc, neighborhood, area, within walk, and locat great all stand out and stress the importance of the location of the property. The figure below represents the WordCloud for the summer 2022 reviews:

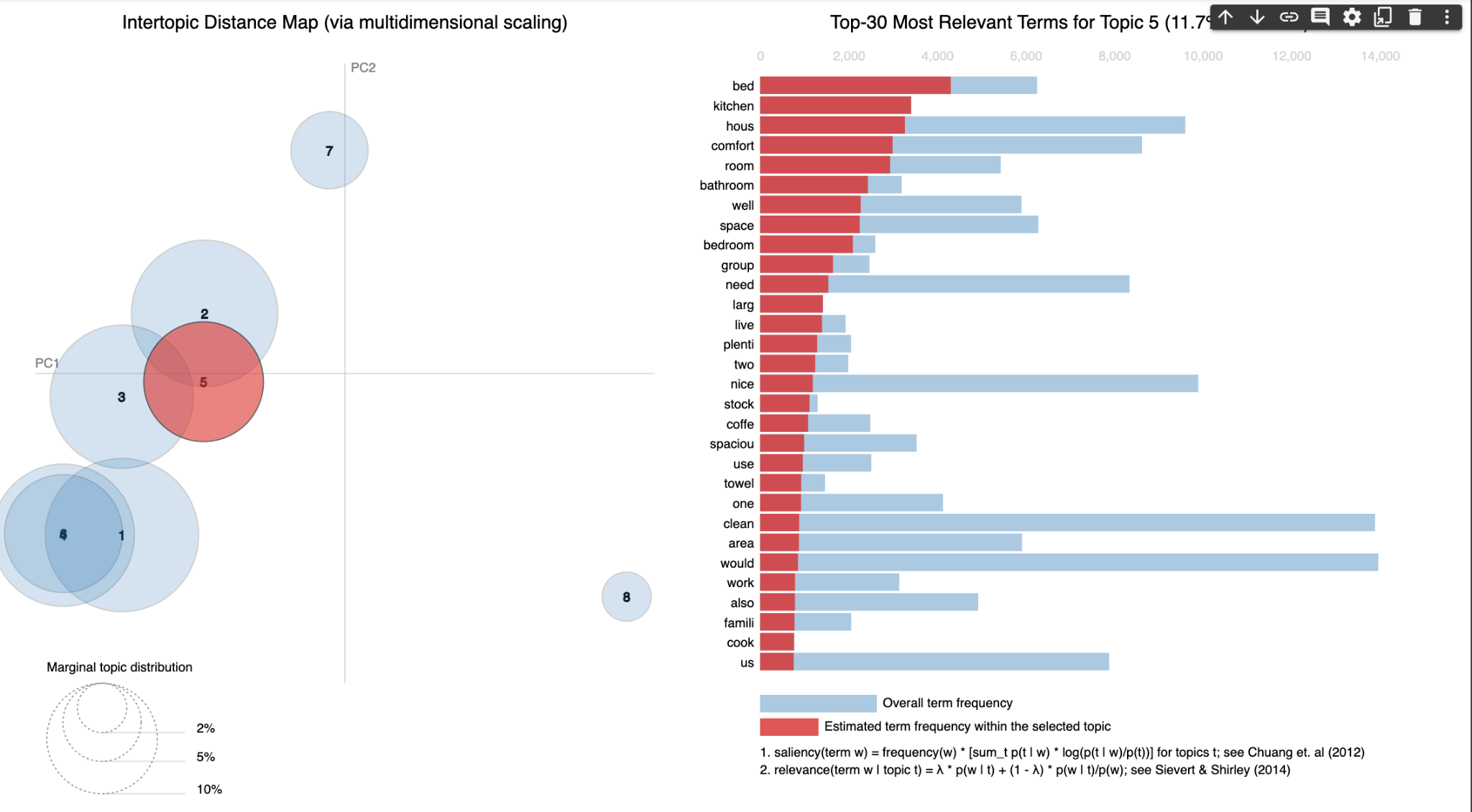


Once again, french quarter, great locat, and walk distanc stand out as terms frequently used relating to location. These two figures are not the end all be all, but from them we are able to begin to infer what features reviewers find important when evaluating a property.

The next section explored for text mining pertains to topic modeling. Topic modeling was completed for both of the reviews datasets, 2023 reviews and summer 2022 reviews. For each review punctuation was removed, all words were converted to lowercase, stopwords were removed, and stemming was performed. This produced a cleaned version of the text data allowing for better analysis when performing topic modeling. The corpus used contained a collection of the reviews for each time period. Once the corpus was created, the next step was to create a dictionary using all the terms inside the corpus. After creating the dictionary, the next step was to create a document-feature matrix using bag of words. And finally using that document-feature matrix we were able to create an LDA model that consisted of 8 topics. The following figure represents the first topic in the 2023 review data:

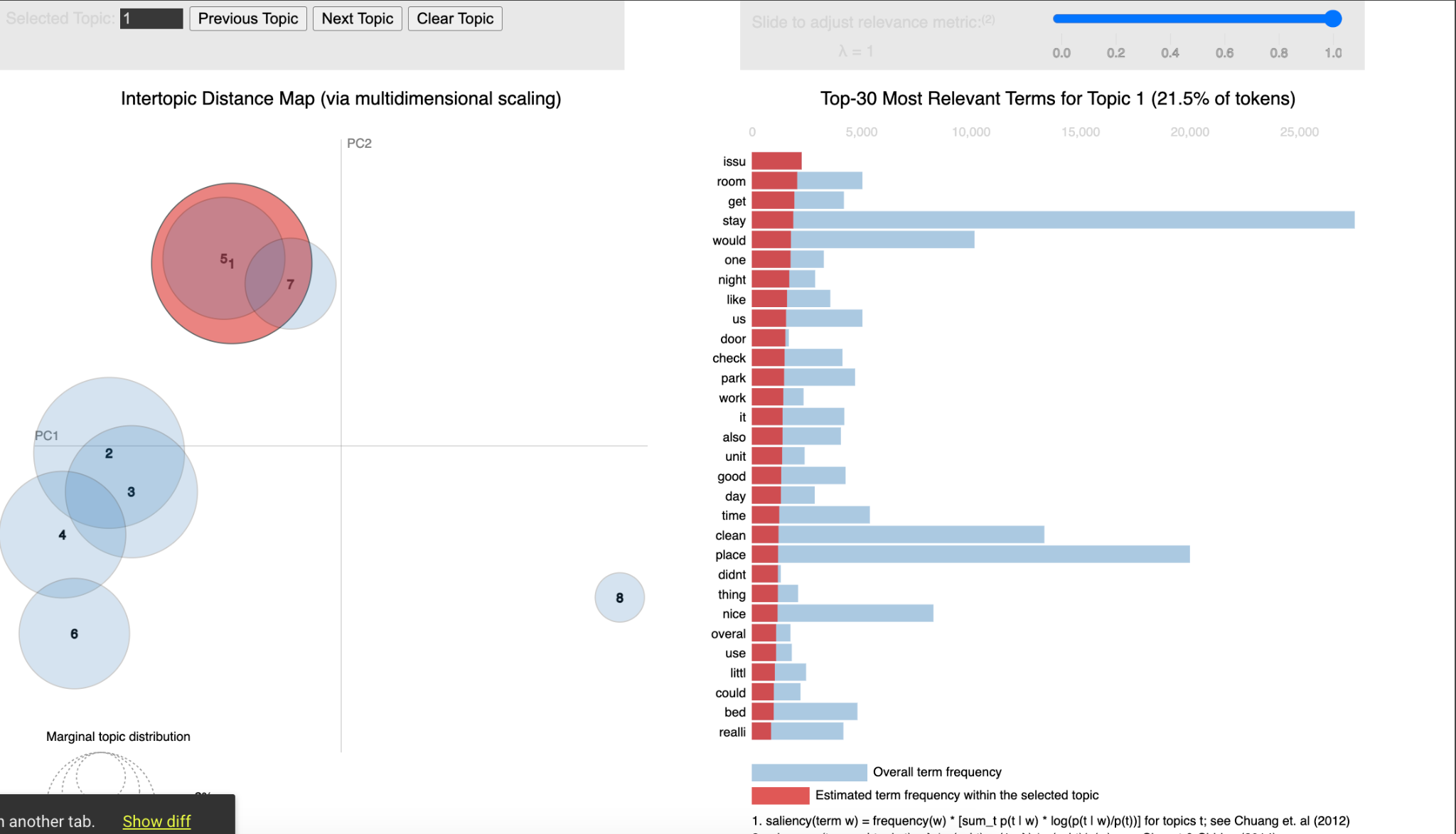


The next figure represents the fifth topic from the 2023 review data:



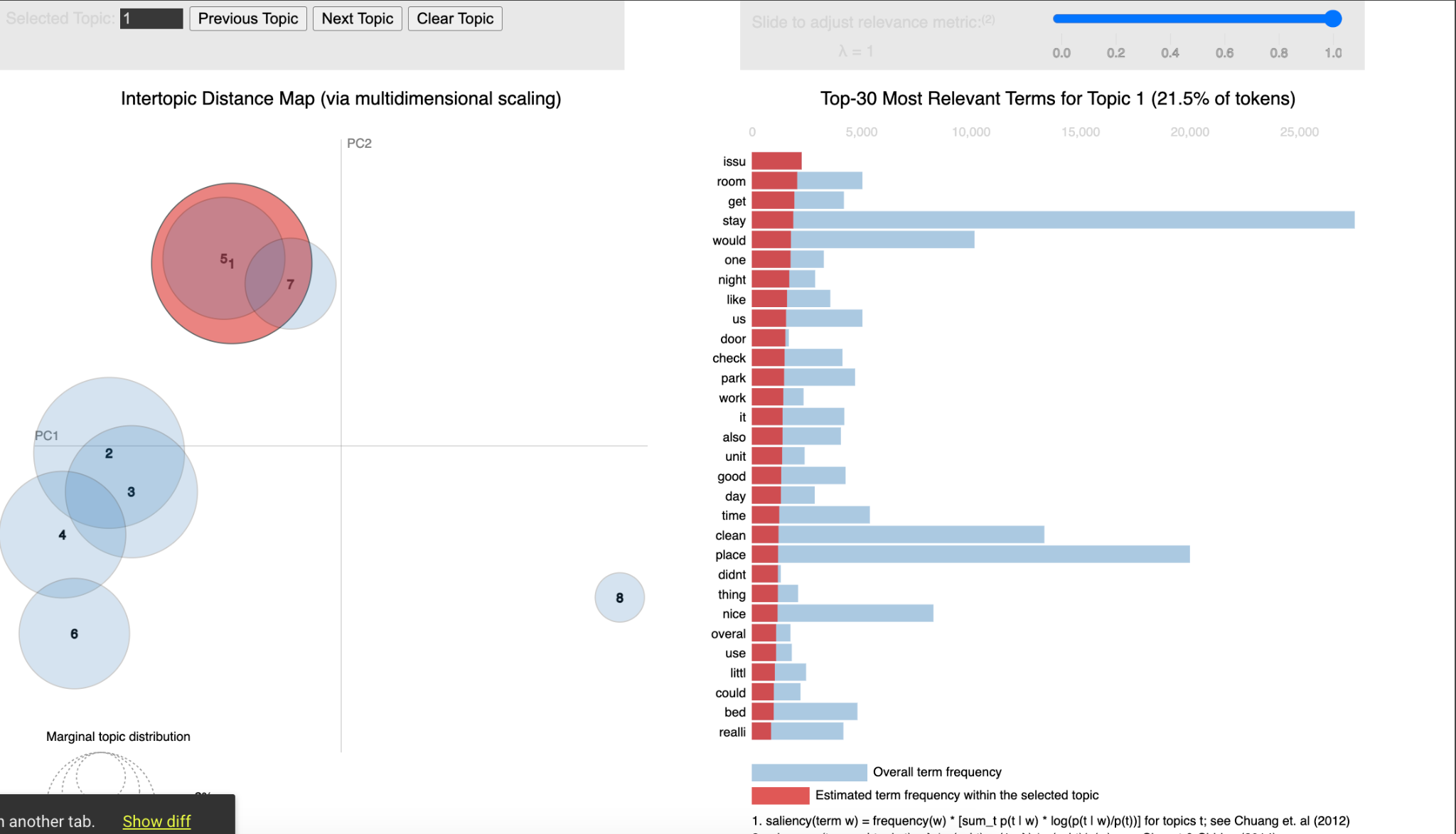
These two topics seem to highlight two areas: the first being location and the second being amenities/features of the property. This gives a nice idea as to what reviewers are looking for when they stay at a property. Using this information, an investor would be able to tailor their property based upon review data, thus resulting in a better average review score.

The next figure is the first topic for the summer 2022 review data:



This topic is difficult to follow and should not be focused upon too much. Sometimes you get a topic that does not provide useful information, which is believed to be the case here.

And the following figure shows the third topic from the summer 2022 data:



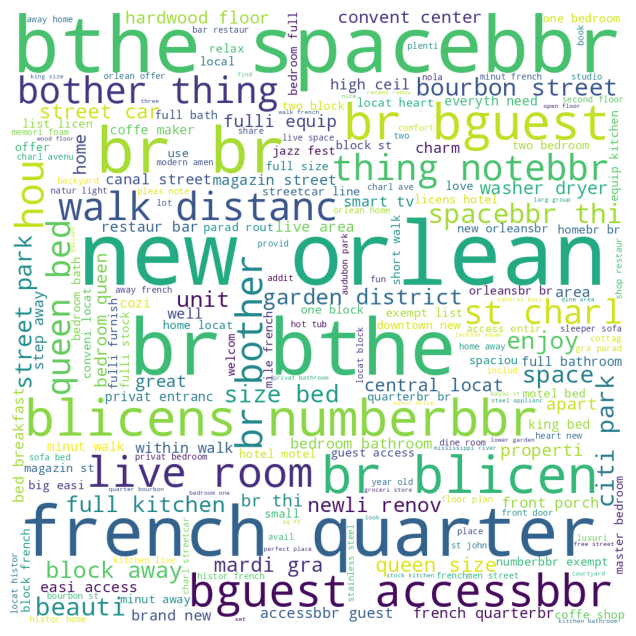
However this topic does seem to have a common theme which is location. Location based text has come up often throughout the text mining process so it is no surprise to see it here again. From this topic you can see the different terms describing the location of the property, like walk, street, french quarter, close, locat, restaur, etc.

In conclusion, the review data seems to mention location, amenities, and defining features of the properties. Therefore it is important to own a property in a good location that offers many amenities.

**Listing Description Text Mining**

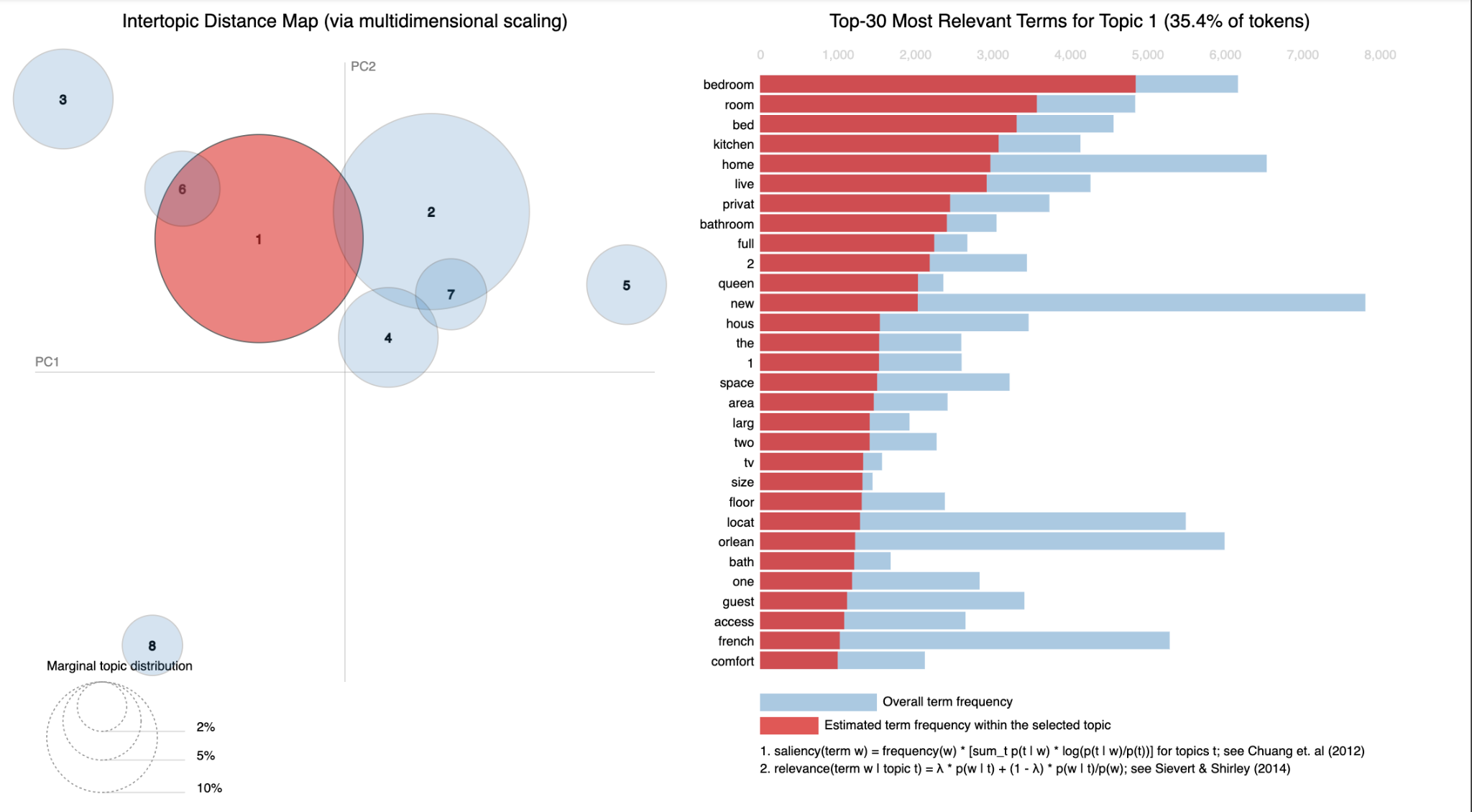
Within our data we had access to a dataset that contained data about each individual property listing. And within that dataset there was a description column where the owner of the property described the unit that was being listed. Using that data we conducted similar research outlined above. The goal here was to reveal commonalities between listings to gain an understanding of what properties have to offer. As an investor, you can use this data to gather ideas as to what may make your property stand out from the rest of the competition. Or also use it to take ideas/inspiration from other properties and incorporate them into yours.

The first step once again was to visualize the data by using a WordCloud. You can see the results below:

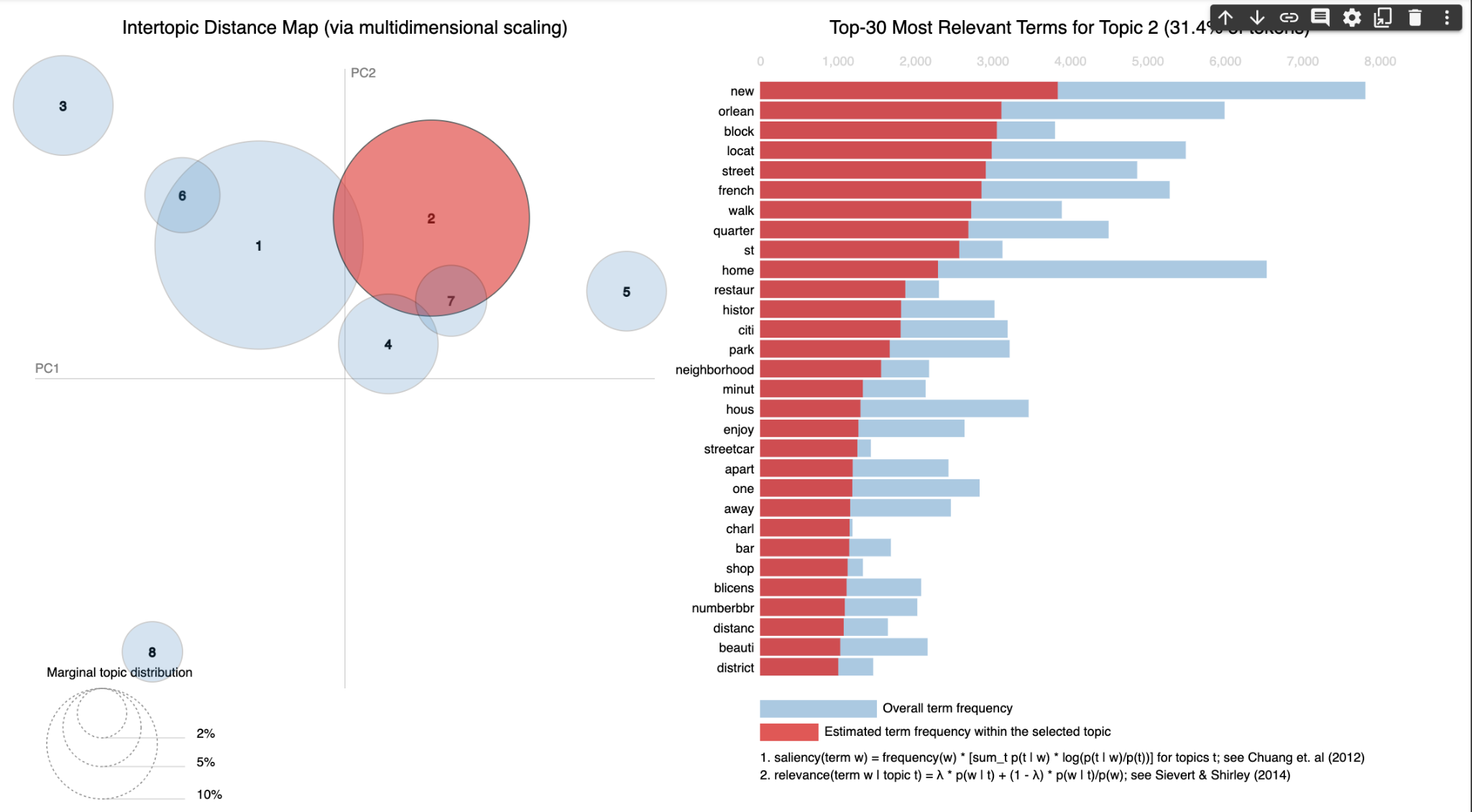


What stands out here is two things: location and amenities. Based upon the WordCloud, it looks as if many descriptions include amenities of the unit such as bedrooms, bathrooms, full kitchen, parking, hot tub, smart tv, etc. And there also seems to be a focus on location with bourbon street, french quarter, and garden district showing up. Along with terms like walk distanc, street car, and block away. One can assume that it is important to describe location and features of the property when creating a description for the listing.

Topic modeling was also conducted for the listing description text. The first topic can be seen below in the figure:



This topic looks to describe the unit listed. Describing the beds, bed sizes, baths, kitchen situation, and defining the features of the space in general. The second topic looks to emphasize location:



Based upon the WordCould and these two topics, it can be concluded that location and property details/amenities is important to include in the listing description.