**Data Preprocessing and Feature Extraction:**

For exploratory data analysis, predictive modelling and trend analysis, first we performed data cleaning over our four main types of Airbnb data sources: listings.csv, neighborhood csv, calendar csv and review csv for Toronto over the years 2015-2019. For seasonality analysis the calendar data was collected in specific date, month, weekdays format, the daily prices are transformed from string to float data and categorical feature: ‘availability’ was shaped into logical value. Major cleaning operations were needed to bring the features of “listing.csv” to specific format for fitting them into specific predictive models. The file contained 52 features, but for per night price prediction most important 20 predictors were selected by running random forest and gradient boosting models. The missing values were imputed and outliers were removed using scikit learning, currency data was formatted into float values and review score range, amenities, response time and other 7 categorical values were encoded differently. For the review rating predictions ‘comments’ from the ‘review.csv’ was processed with natural language processing library of python. ‘neighbourhood.csv’ was used for neghbourhood modeling by processing the “Zipcode” and “Neighbourhood’ variables and extracting latitude and longitude from them.

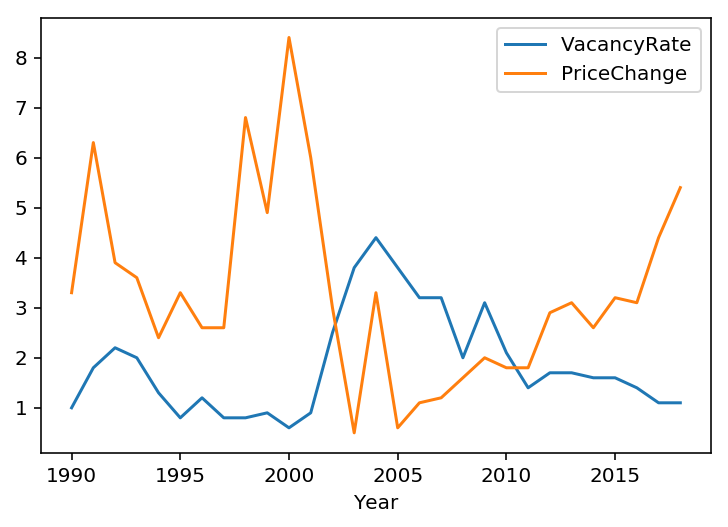
1. Trend and Seasonality Analysis of Airbnb Price:

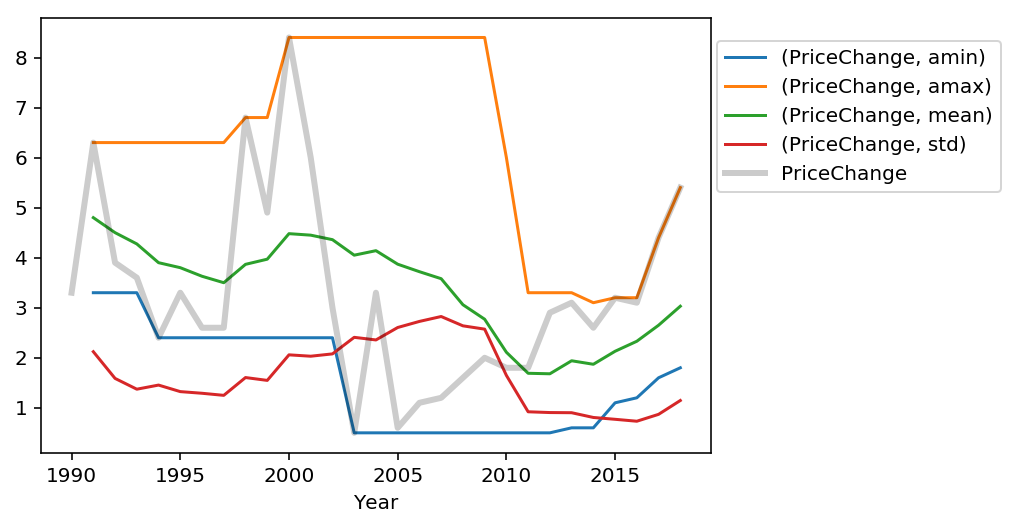
The main files for data modeling: data-modeling-baseline.ipynb, data-modeling-baseline-cluster.ipynb, seasonality-exploration-WJ.ipynb. Various regression models were fit such as Linear Regression, Ridge Regression, and Lasso Regression. While these models seemed to indicate a poor fit based on their R2R2 score (~0.3 for a test dataset score), analyzing by Median Absolute Error proved to be significantly more insightful. Evaluating by Median Absolute Error is a more robust method than RSS in this setting, as it is less affected by outliers (an issue that was raised in predicting pricing especially for Airbnb listings) and yields a more intuitive score that is easily translatable to pricing. A Ridge Regression obtained an optimal Median Absolute Error of $21.43. However, when analyzing by single house listing data, a Bayes Ridge Regression obtained an optimal Median Absolute Error of $19.43 (Figure 2).

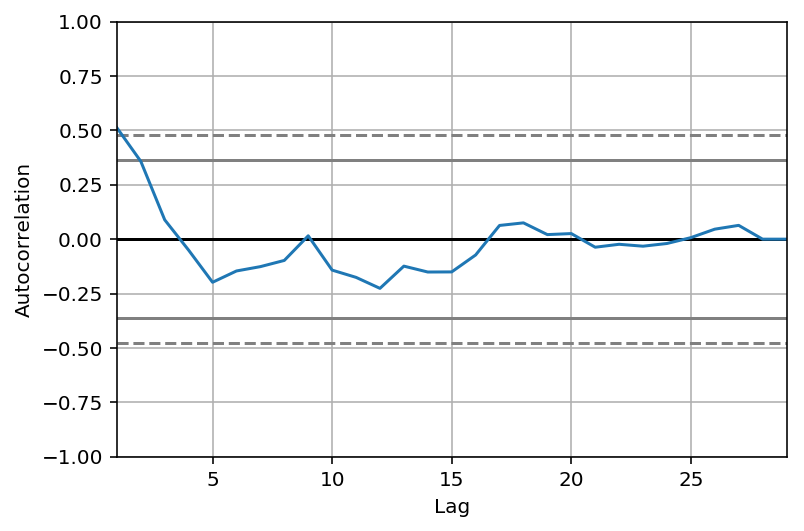
#### Incorporating Seasonality

After cleaning the seasonality data, the main source of analysis on the seasonality data is in the seasonality-exploration-WJ.ipynb. Overall, looking at seasonality was a worthwhile endeavor. There were noticeable changes in the prices depending on the day of the week. On average, Friday and Saturday listings proved to have the highest price on Airbnb. Both are priced on average 103% the Monday price. Tuesday and Wednesday are on average the least expensive, both around 99% monday prices. Overall, the results suffered from greater error than our original RidgeCV regression, as we experienced a median absolute error of $72.10. This is for three reasons -- many of the listings still do not incorporate day-variations in their prices, inaccuracies within the original RidgeCV predictions, and large rather than small price difference between days of the week. However, this area perhaps is the most promising because it shows that many Airbnb listers are not taking advantage of dynamic pricing by the day of the week, something that is important to establish optimal pricing. Already, there are some promising results -- there is evidence to suggest that hosts should price Friday and Saturday the highest and Tuesday and Wednesday the lowest. With further refinements to our model, such as looking at various seasonal time series models, people may be able to look at the trends and price their Airbnb listings more appropriately and with more concrete percentage change suggestions.

Trend Analysis and Prediction For Long Term Rental:





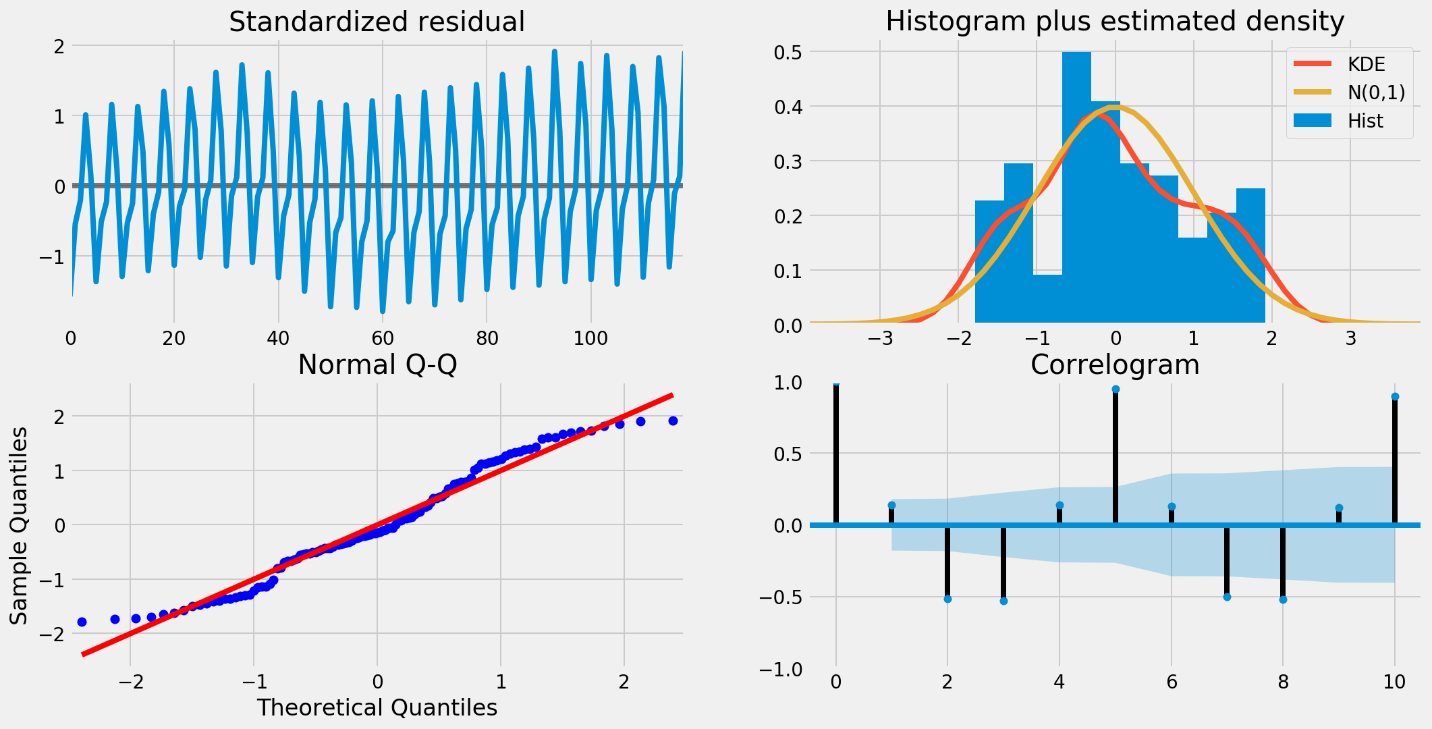


When we plot the autocorrelation of the 'PriceChange' series: on the x-axis, you have the lag and, on the y-axis, you have how correlated the time series is with itself at that lag. So, this means that if the original time series repeats itself every two years, you would expect to see a spike in the autocorrelation function at 2 years. Looking at the plot we see a spike in the autocorrelation function at 9 years: the time series is correlated with itself shifted by 9 years. So, we have identified the seasonality of 9 years repetition!

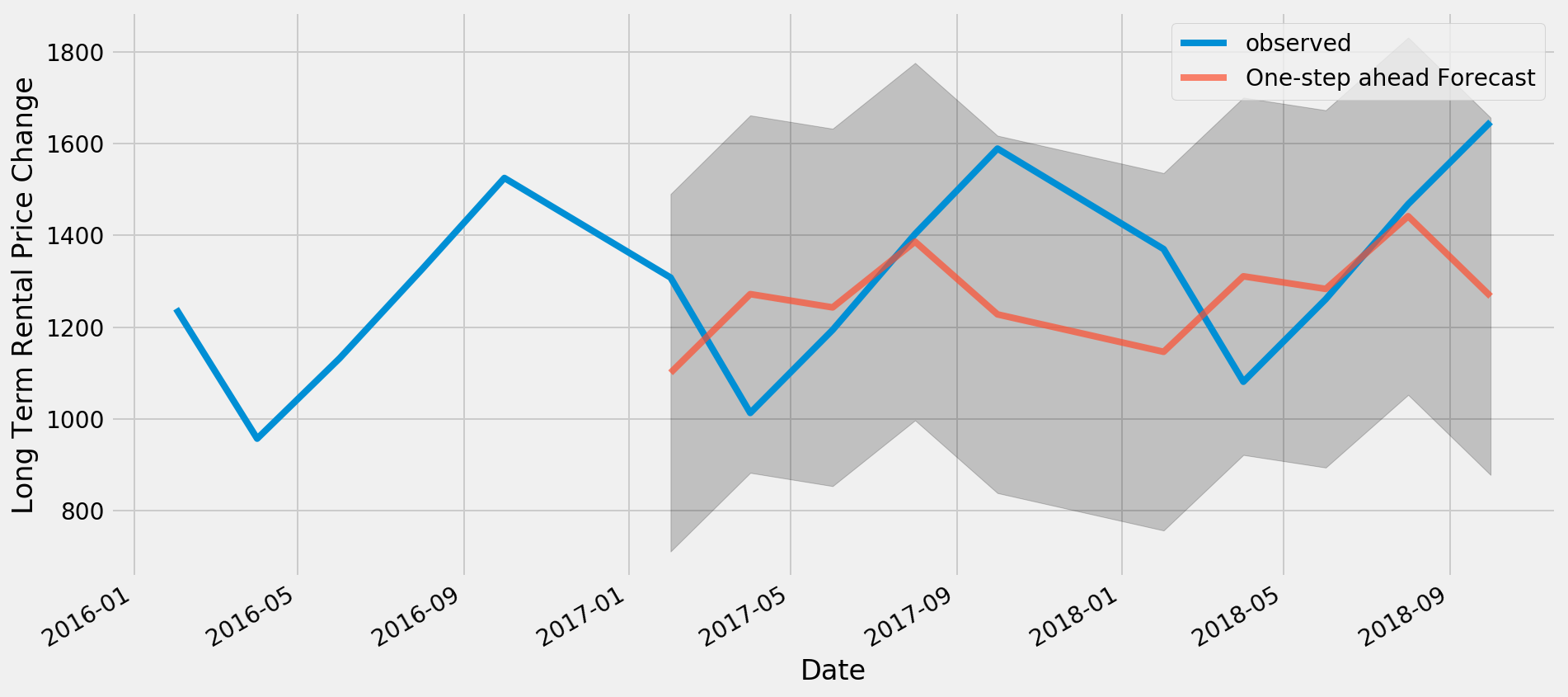
**Time series forecasting with ARIMA**

We are going to apply one of the most commonly used method for time-series forecasting, known as ARIMA, which stands for Autoregressive Integrated Moving Average.

ARIMA models are denoted with the notation ARIMA(p, d, q). We should always run model diagnostics to investigate any unusual behavior. It is not perfect, however, our model diagnostics suggests that the model residuals are near normally distributed.



To help us understand the accuracy of our forecasts, we compare predicted sales to real sales of the time series, and we set forecasts to start at 2017–02–01 to the end of the data.



The Mean Squared Error of our forecasts is 1.16. The line plot is showing the observed values compared to the rolling forecast predictions. Overall, our forecasts align with the true values very well, showing an upward trend starts from the beginning of the year and captured the seasonality toward the end of the year.

####For Airnb part

##Housing Investment:

Factors: Accomodation Cost, Vacancy Rate, New Migration, Population how influencing new housing construction investment and overall housing market price:

**Using data to understand the market for AirBnB rentals in Seattle**

## **I looked at AirBnB data to see how homeowners could offer a great guest experience and maximize their revenues**

**Introduction**

Before AirBnB, it would have been a nerve wracking prospect to let strangers stay in your home. AirBnB has changed this — with a mission that [“connects people with places to stay and things to do around the world.”](https://www.airbnb.com/host/homes?from_nav=1)

The company has transformed the relationship between the homeowner and the renter. Most of us are familiar with the experience as guests, renting homes to stay in on AirBnB. But I was interested in the perspective of a homeowner.

AirBnB projects the prospect of making money by renting out your home with the platform. But homeowners, especially those renting out their homes for the first time, may have many questions: What price should I set my home at? Can I trust my home to guests? How can I ensure I get a good rating?

This [dataset](https://www.kaggle.com/airbnb/seattle/home), generously provided by AirBnB on Kaggle, covering Seattle listings on the site has the potential to shed some light on these questions. The data covers 3,818 listings on AirBnB in Seattle. It provides information on home features, review scores, and their availability in the year 2016 and the very first days of 2017. We have only two day’s worth of data for 2017, so this analysis is focused on 2016 data.

Ultimately, as a prospective Seattle homeowner, my main objective would be to**offer a great experience for my guests while maximizing revenues**. To this end, I dove into the data.

**Part I: What could be the main factors driving higher ratings?**

The review scoring system on AirBnB is the main way in which guests can leave feedback on their hosts, and vice versa. Hence, as a Seattle homeowner, I would want to investigate what would drive guests to leave a high rating on AirBnB. A high rating would also have a positive feedback loop, in that homes with a higher rating would tend to attract other guests to stay there, ensuring that a home continues generating rentals.

In order to identify the drivers of higher ratings, I used a multiple linear regression model. An explanation of the math behind this model is beyond the scope of this article, but the general idea behind it is to determine the underlying trend of tendency (in math, called the correlation) of features in a home with the rating of that home.

In line with this idea, if the trend is a positive correlation, this means that the higher quantity of a feature of the home that is present, the higher the rating would tend to be. The model also gives the ‘weights’ of each feature that affects the rating. This means that the higher the weight of the feature, the bigger the effect of that feature on the overall rating.

Ultimately, what I discovered within the data was that the correlation was only **slightly positive**, meaning that I cannot say with certainty that factors present in the data are all of the drivers that lead to a higher rating. The relationship between the factors and that they lead to a higher rating is thus **inconclusive**.

Meanwhile, the top five factors—according to the weights that the model discovered are the ‘strongest’ factors (meaning that they have a higher tendency to lead to higher ratings)—are as follows:

1. **Count of host listings**: the number of listings a host has

2. **Review scores —**[**value**](https://www.airbnb.com/help/article/1257/how-do-star-ratings-work): did the guest feel the listing provided good value for the price?

3. **Review scores —**[**cleanliness**](https://www.airbnb.com/help/article/1257/how-do-star-ratings-work): did the guests feel that the space was clean and tidy?

4. **Review scores —**[**accuracy**](https://www.airbnb.com/help/article/1257/how-do-star-ratings-work): how accurately did the listing page represent the space?

5.**Review scores —**[**communication**](https://www.airbnb.com/help/article/1257/how-do-star-ratings-work): how well did the host communicate with the guest before and during their stay?

The highest weighted feature—the count of host listings—could be a proxy indicator for credible and experienced hosts. It would therefore not be a surprise that more experienced hosts tend to be more knowledgeable in what guests would appreciate in a home, hence leading to higher ratings.

Meanwhile, the other four features constitute the review scores of other features of the home which guests are allowed to review (besides providing an overall rating) on AirBnB. It is no surprise that the model finds that guests appreciate good value for money when it comes to their stay.

However, it could be actionable to reduce the price if previous guests have been giving lower scores, as it indicates that the home might not be worth the price set. Some other actions could be to add additional amenities or extra services to the home to increase the guest’s perception of the value the home represents.

Meanwhile, a good rating is correlated with high cleanliness, as well as accuracy — hence, as a host, I would ensure my homes are clean and that the pictures and description are truthful and reflective of the home.

Lastly, good communication is weighted with the tendency to give a high rating. Hence, a Seattle homeowner should ensure they are responsive to guests, as this is a factor that leads to the guests’ rating an overall highly positive experience.

It is worth noting that the above features are weighted according to the same model that states the correlation is relatively weak, even if it is positive. However, even if the relationship is not conclusive, the factors above are still actionable, good practices that a prospective Seattle AirBnB host should incorporate when renting out their home on AirBnB. Not practicing them may hurt their overall rating.

**Conclusion**

In this article, we explored how Seattle homeowners might best position themselves to provide a great experience for guests, while maximizing revenues from renting out their homes on AirBnB. We used data from AirBnB on Seattle listings in 2016 to the first two days of 2017 to achieve this objective, with the following findings:

1. We attempted to discover if we can predict the features of homes that lead guests to give higher overall ratings. We found that while **the relationship between the features of a particular home and the overall rating is only slightly positively correlated**, there could be merit in trying to signal that you are a **credible and experienced host**, while having providing **good value for the price of renting a home**, **high cleanliness**, **accurate descriptions and pictures** of the home, and **responsive communication** with guests.
2. We then explored the most popular times of the year for Seattle home rentals in 2016, with very limited data for 2017. We found that**January 2016 had the highest number of home rentals**, but this number **rapidly decreased as the year progressed** and even extended into the first two days 2017.
3. Finally, we looked at the neighbourhoods and times of the year where average prices of listings were maximized, in order to determine the maximum revenue generating opportunities of a homeowner. We found that the **Broadway neighbourhood generated the highest overall revenues**, due to its popularity and average price combination. **March 2016 provided the highest average price of a listing in the year 2016**, with no data for 2017.

# Making Models (I) | Airbnb Price Prediction: Data Analysis

## **Location, location, … cancellation? What causes Airbnb price difference?**

[Go to the profile of Philip Mohun](https://medium.com/@philmohun?source=post_header_lockup)

[Philip Mohun](https://medium.com/@philmohun)Follow

Feb 27, 2018

This is part one of a series documenting the end to end process to develop and design a generalized linear model that outputs predicted Airbnb rental price. As a whole, the series will include a **description of dataset analysis**, [advice for additional data collection via web scraping methods](https://medium.com/@philmohun/making-models-airbnb-price-prediction-data-collection-via-web-scraping-6218f35cdebb), [feature engineering (specifically for unstructured images)](https://medium.com/@philmohun/making-models-airbnb-price-prediction-feature-engineering-and-unstructured-image-analysis-8f0456663fd8), model selection and results. Readers should find this document helpful when developing their own predictive models, or when looking for a framework to organize their thoughts. Most importantly, I hope to demystify some of the processes behind “data science” by breaking down a typical workflow into distinct and modular activities that can be reproduced for many types of problems. If you find it helpful or have a question, please feel free to leave a comment below and I will answer to the best of my ability.

Today we will break our analysis down into several parts, each with a specific goal in mind. These include:

1. **Description of dataset characteristics**and explanation of each variable in plain English.
2. **Target variable examination**to gain an understanding of possible influences.
3. **Multivariate study** of each feature, and potential relations between variables.

By the end of this section, we should have a better understanding of the features that make up our dataset and how they impact our target variable.

#### *Dataset Description*

To accurately predict Airbnb price, we aim to collect a dataset containing features which directly impact the rental price. No better place to start than by gathering a number of listings with fields directly from the site. Below you will find a list of the features that were taken from Airbnb and which turn out to be very important attributes in the price prediction. Since we know the price for each row, this can be classified as a supervised learning problem, and we will split our data into distinct training, test, and cross-validation sets. For now, we will examine the dataset as a whole, and come back to this division later. As a general rule, I like to examine a dataset’s features for several characteristics before proceeding or deciding to gather additional data. These characteristics include:

* Number of missing values and how to deal with them (NaN or null)
* Type of data (categorical, boolean, image, numerical, text, etc)
* Shape and size of data (this impacts the type of model we will use)
* Classical statistical analysis (mean, median, range, variance, st. dev, etc)

**Understanding the problem**

At a glance:

**Target variable:** Log price (natural logarithm)

**Features of dataset:**

* id (numeric) | Unique identifier for each listing
* property\_type (categorical) | (e.g. Apartment, house, condo)
* room\_type (categorical) | (e.g. Entire home/apt, private room)
* amenities (text) | Unstructured list separated by commas (e.g. tv, kitchen). Candidate for textual analysis.
* accommodates (numeric) | Number of people the rental fits
* bathrooms (numeric) | Number of full and/or half baths
* bed\_type (categorical) | (e.g. futon, real bed)
* cancellation\_policy (categorical) | (e.g. Flexible, moderate, strict)
* cleaning\_fee (boolean) | T/F
* city (categorical) | (e.g. Boston, NYC, LA)
* description (text) | Unstructured and up to the host how to populate. Candidate for textual analysis.
* first\_review (date) | How long ago the first review was left
* host\_has\_profile\_pic (boolean) | T/F (no link to picture)
* host\_identity\_verified (boolean) | T/F (via email verification)
* host\_response\_rate (numeric) | How often the host replies to inquiries (%)
* host\_since (date) | Date that they opened their account
* instant\_bookable (boolean) | T/F
* last\_review (date) | Date of most recent hosting
* latitude (numeric)
* longitude (numeric)
* name (text) | Name of rental property. Candidate for textual analysis.
* neighbourhood (categorical) | Informal description of neighborhood (e.g. Brooklyn Heights, Downtown)
* number\_of\_reviews (numeric) | Total number of reviews given by guests
* review\_scores\_rating (numeric) | Mean rating of reviews given by guests
* thumbnail\_url (numeric; we’ll come back to this) | Link to primary photo of rental property. Candidate for image analysis.
* zipcode (numeric; likewise) | Zipcode. Candidate for bringing in additional data.
* bedrooms (numeric) | Number of bedrooms in rental
* beds (numeric) | Number of beds in rental

Let’s break this down a bit.

The type of feature (categorical, numeric, boolean, text) will impact the way that we perform our analysis and choose our eventual model. For our numeric indicators we can perform statistical analysis, but for our categorical and text data, we will have to get a bit more creative.

The purpose of this portion was to look at our current information and get a feel for what we are working with. The Airbnb website helpfully provided these features, but it does not tell us which is more or less important as an indicator of price. Based on anecdotal experience (a.k.a. staying in Airbnbs), I would guess that the number of beds, accommodates, and zip code are probably important for determining the price. For now, I’m going to leave all features intact — we will return to this during our multivariate analysis!

#### Target variable analysis

Recall that we are attempting to predict log\_price as our target variable. I’ll be using Python with Jupyter notebooks to do some of the manipulations and will include code snippets when applicable.

**Data Cleaning Parts**

[Homepage](https://medium.com/)

[Data Driven Investor](https://medium.com/datadriveninvestor?source=logo-8bcefc526e02---32881626c9c9)

Follow

Fahmida Brishty

* [ABOUT](https://medium.com/datadriveninvestor/about)
* [VIDEOS](https://medium.com/datadriveninvestor/videos/home)
* [COLLABORATE](https://medium.com/datadriveninvestor/write-for-data-driven-investor-1860db5565a4)
* [SUBSCRIBE](https://medium.com/datadriveninvestor/subscribe-to-our-newsletter-to-recieve-our-weekly-updates-more-straight-to-your-inbox-aeaf40f4c5b6)
* [AI](https://medium.com/datadriveninvestor/tagged/artificial-intelligence)
* [DDI](http://www.datadriveninvestor.com/)

[https://cdn-images-1.medium.com/max/800/0*bE6SIr5J8WyH1aqh](http://www.track.datadriveninvestor.com/1B9E)

# Making Models (I) | Airbnb Price Prediction: Data Analysis

## **Location, location, … cancellation? What causes Airbnb price difference?**

[Go to the profile of Philip Mohun](https://medium.com/@philmohun?source=post_header_lockup)

[Philip Mohun](https://medium.com/@philmohun)Follow

Feb 27, 2018

This is part one of a series documenting the end to end process to develop and design a generalized linear model that outputs predicted Airbnb rental price. As a whole, the series will include a **description of dataset analysis**, [advice for additional data collection via web scraping methods](https://medium.com/@philmohun/making-models-airbnb-price-prediction-data-collection-via-web-scraping-6218f35cdebb), [feature engineering (specifically for unstructured images)](https://medium.com/@philmohun/making-models-airbnb-price-prediction-feature-engineering-and-unstructured-image-analysis-8f0456663fd8), model selection and results. Readers should find this document helpful when developing their own predictive models, or when looking for a framework to organize their thoughts. Most importantly, I hope to demystify some of the processes behind “data science” by breaking down a typical workflow into distinct and modular activities that can be reproduced for many types of problems. If you find it helpful or have a question, please feel free to leave a comment below and I will answer to the best of my ability.

Today we will break our analysis down into several parts, each with a specific goal in mind. These include:

1. **Description of dataset characteristics**and explanation of each variable in plain English.
2. **Target variable examination**to gain an understanding of possible influences.
3. **Multivariate study** of each feature, and potential relations between variables.

By the end of this section, we should have a better understanding of the features that make up our dataset and how they impact our target variable.

#### *Dataset Description*

To accurately predict Airbnb price, we aim to collect a dataset containing features which directly impact the rental price. No better place to start than by gathering a number of listings with fields directly from the site. Below you will find a list of the features that were taken from Airbnb and which turn out to be very important attributes in the price prediction. Since we know the price for each row, this can be classified as a supervised learning problem, and we will split our data into distinct training, test, and cross-validation sets. For now, we will examine the dataset as a whole, and come back to this division later. As a general rule, I like to examine a dataset’s features for several characteristics before proceeding or deciding to gather additional data. These characteristics include:

* Number of missing values and how to deal with them (NaN or null)
* Type of data (categorical, boolean, image, numerical, text, etc)
* Shape and size of data (this impacts the type of model we will use)
* Classical statistical analysis (mean, median, range, variance, st. dev, etc)

**Understanding the problem**

At a glance:

**Target variable:** Log price (natural logarithm)

**Features of dataset:**

* id (numeric) | Unique identifier for each listing
* property\_type (categorical) | (e.g. Apartment, house, condo)
* room\_type (categorical) | (e.g. Entire home/apt, private room)
* amenities (text) | Unstructured list separated by commas (e.g. tv, kitchen). Candidate for textual analysis.
* accommodates (numeric) | Number of people the rental fits
* bathrooms (numeric) | Number of full and/or half baths
* bed\_type (categorical) | (e.g. futon, real bed)
* cancellation\_policy (categorical) | (e.g. Flexible, moderate, strict)
* cleaning\_fee (boolean) | T/F
* city (categorical) | (e.g. Boston, NYC, LA)
* description (text) | Unstructured and up to the host how to populate. Candidate for textual analysis.
* first\_review (date) | How long ago the first review was left
* host\_has\_profile\_pic (boolean) | T/F (no link to picture)
* host\_identity\_verified (boolean) | T/F (via email verification)
* host\_response\_rate (numeric) | How often the host replies to inquiries (%)
* host\_since (date) | Date that they opened their account
* instant\_bookable (boolean) | T/F
* last\_review (date) | Date of most recent hosting
* latitude (numeric)
* longitude (numeric)
* name (text) | Name of rental property. Candidate for textual analysis.
* neighbourhood (categorical) | Informal description of neighborhood (e.g. Brooklyn Heights, Downtown)
* number\_of\_reviews (numeric) | Total number of reviews given by guests
* review\_scores\_rating (numeric) | Mean rating of reviews given by guests
* thumbnail\_url (numeric; we’ll come back to this) | Link to primary photo of rental property. Candidate for image analysis.
* zipcode (numeric; likewise) | Zipcode. Candidate for bringing in additional data.
* bedrooms (numeric) | Number of bedrooms in rental
* beds (numeric) | Number of beds in rental

Let’s break this down a bit.

The type of feature (categorical, numeric, boolean, text) will impact the way that we perform our analysis and choose our eventual model. For our numeric indicators we can perform statistical analysis, but for our categorical and text data, we will have to get a bit more creative.

The purpose of this portion was to look at our current information and get a feel for what we are working with. The Airbnb website helpfully provided these features, but it does not tell us which is more or less important as an indicator of price. Based on anecdotal experience (a.k.a. staying in Airbnbs), I would guess that the number of beds, accommodates, and zip code are probably important for determining the price. For now, I’m going to leave all features intact — we will return to this during our multivariate analysis!

#### Target variable analysis

Recall that we are attempting to predict log\_price as our target variable. I’ll be using Python with Jupyter notebooks to do some of the manipulations and will include code snippets when applicable.

count 19255.000000

mean 139.749935

std 210.438228

min 0.000000

25% 65.000000

50% 100.000000

75% 158.000000

max 13315.000000

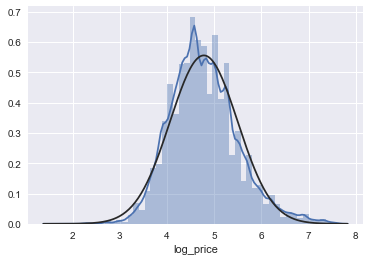
Name: price, dtype: float64

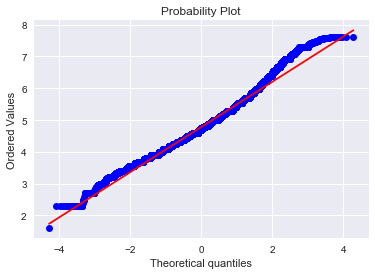
We can see that we’ve gathered ~19255 rows of information, and log\_price fluctuates between 0 and 7.7. The zero value here is problematic and will require a closer examination. If hosts are offering free rentals I’d like to know!

Only one listing had a “0” value for log\_price — I knew it was too good to be true.

We’ve removed the row with a zero value for log\_price — much better.

Getting an idea of this via a histogram will help to determine if any additional transformations need to be made before proceeding.





Minimally skewed and showing insignificant kurtosis (<1) this passes the spot check for normal distribution (bell-shaped, symmetrical about the center). Likewise, our probability plot appears linear and reinforces our decision to leave log\_price as-is.

We now have painted ourselves a picture of log\_price that will be useful when examining the results of our model. When the time comes to predict results we can compare our test set to this distribution and determine if we are in the ballpark. We still have some work to do before we get there — let’s dive into what causes it to tick.

#### Multivariate study

Taking a look at how this set stacks up we already know that we have gathered ~74,000 rows, giving us some leeway if we decide to remove listings lacking information. In fact, it’s probably a good idea to see which are the biggest culprits for missing values.

id 19255 non-null int64

listing\_url 19255 non-null object

scrape\_id 19255 non-null int64

last\_scraped 19255 non-null int64

name 19255 non-null object

summary 18622 non-null object

space 13467 non-null object

description 18944 non-null object

experiences\_offered 19255 non-null object

neighborhood\_overview 11990 non-null object

notes 9081 non-null object

transit 12238 non-null object

access 12092 non-null object

interaction 11495 non-null object

house\_rules 12882 non-null object

thumbnail\_url 0 non-null float64

medium\_url 0 non-null float64

picture\_url 19255 non-null object

xl\_picture\_url 0 non-null float64

host\_id 19255 non-null int64

host\_url 19255 non-null object

host\_name 19252 non-null object

host\_since 19252 non-null float64

host\_location 19240 non-null object

host\_about 11269 non-null object

host\_response\_time 10901 non-null object

host\_response\_rate 10901 non-null float64

host\_acceptance\_rate 0 non-null float64

host\_is\_superhost 19252 non-null object

host\_thumbnail\_url 19252 non-null object

host\_picture\_url 19252 non-null object

host\_neighbourhood 16674 non-null object

host\_listings\_count 19252 non-null float64

host\_total\_listings\_count 19252 non-null float64

host\_verifications 19255 non-null object

host\_has\_profile\_pic 19252 non-null object

host\_identity\_verified 19252 non-null object

street 19255 non-null object

neighbourhood 18161 non-null object

neighbourhood\_cleansed 19255 non-null object

neighbourhood\_group\_cleansed 0 non-null float64

city 19254 non-null object

state 19244 non-null object

zipcode 18836 non-null object

market 19213 non-null object

smart\_location 19255 non-null object

country\_code 19255 non-null object

country 19255 non-null object

latitude 19255 non-null float64

longitude 19255 non-null float64

is\_location\_exact 19255 non-null object

property\_type 19255 non-null object

room\_type 19255 non-null object

accommodates 19255 non-null int64

bathrooms 19242 non-null float64

bedrooms 19248 non-null float64

beds 19236 non-null float64

bed\_type 19255 non-null object

amenities 19255 non-null object

square\_feet 167 non-null float64

price 19255 non-null int64

weekly\_price 2321 non-null float64

monthly\_price 1999 non-null float64

security\_deposit 14353 non-null float64

cleaning\_fee 15877 non-null float64

guests\_included 19255 non-null int64

extra\_people 19255 non-null int64

minimum\_nights 19255 non-null int64

maximum\_nights 19255 non-null int64

calendar\_updated 19255 non-null object

has\_availability 19255 non-null object

availability\_30 19255 non-null int64

availability\_60 19255 non-null int64

availability\_90 19255 non-null int64

availability\_365 19255 non-null int64

calendar\_last\_scraped 19255 non-null int64

number\_of\_reviews 19255 non-null int64

first\_review 15526 non-null float64

last\_review 15527 non-null float64

review\_scores\_rating 15262 non-null float64

review\_scores\_accuracy 15250 non-null float64

review\_scores\_cleanliness 15252 non-null float64

review\_scores\_checkin 15247 non-null float64

review\_scores\_communication 15253 non-null float64

review\_scores\_location 15244 non-null float64

review\_scores\_value 15246 non-null float64

requires\_license 19255 non-null object

license 0 non-null float64

jurisdiction\_names 2 non-null object

instant\_bookable 19255 non-null object

is\_business\_travel\_ready 19255 non-null object

cancellation\_policy 19255 non-null object

require\_guest\_profile\_picture 19255 non-null object

require\_guest\_phone\_verification 19255 non-null object

calculated\_host\_listings\_count 19255 non-null int64

reviews\_per\_month 15526 non-null float64

Removed Columns:

scrape\_id, last\_scraped, picture\_url, host\_picture\_url,, neighbourhood\_group\_cleansed, square\_feet, monthly\_price, weekly\_price weekly and monthly removed since we are focusing on single day listings, notes, host\_acceptance\_rate, jurisdiction\_names, interaction. square\_feet (96% of original dataset) missing column values - too many to impute. After removal we have 74 columns but see that still they have a lot of missing values. We drop any entries that are missing (NaN) values and convert the format in price from $1.00 into a float of 1.00. We also drop any entries that are inconsistent; i.e. predictors accommodates, bedrooms, beds, or price with a value of 0. Convert Zip Code values such as 10022-4175 into 10022. Finally, we get rows of 13782.

We explored the number of accommodates in the listed Airbnb and found the distribution like this:

Accommodation 1: 2643

Accommodation 2: 11400

Accommodation 3: 2909

Accommodation 4: 4278

Accommodation 5: 982

Accommodation 6: 1214

Accommodation 7: 217

Accommodation 8: 333

Accommodation 9: 57

Accommodation 10: 119

Accommodation 11: 15

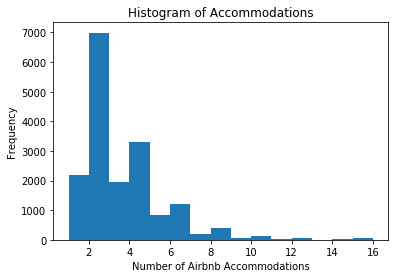
Accommodation 12: 43

Accommodation 13: 4

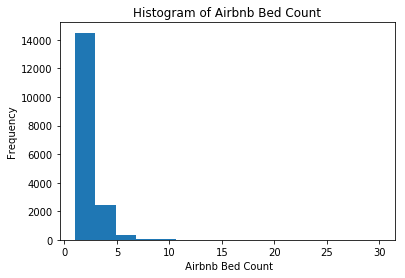
Accommodation 14: 14

Accommodation 15: 5

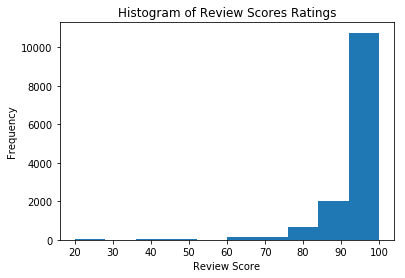
Accommodation 16: 69



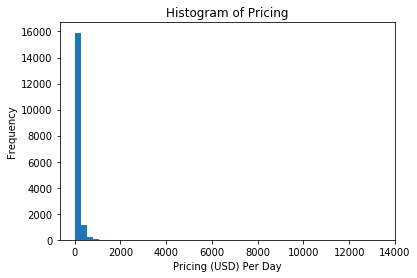
We see that a majority of listings have accomodations for 1-6 people. 1 bed typically accomodates 2 individuals, so let's plot beds instead to analyze how many of the listings are single bedroom listings.



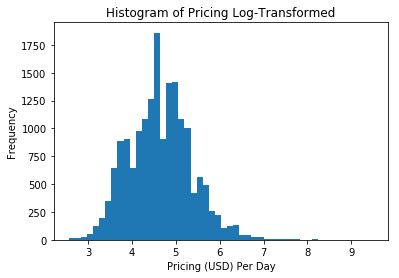
We see that a majority of our listings are indeed single bedroom listings roughly 72%.



11000 listings get review score in between 90 to 100. here are 7712 listings with no review\_scores\_rating (NaN) values. We will handle this by turning the review\_scores\_ratings into a categorical predictor. Based on the histogram, we see that a vast majority of the listings seem to have overall favorable ratings; i.e. skewed left distribution. We replace any NaN values that have no review with 'No Reviews' and we remove any remaining inconsistent NaN values that have a number\_of\_reviews > 0. We will also convert the review\_scores\_ratings into buckets. We will have the bucket ranges set more closely to a higher review score to accomodate for the skewed left distribution of the review scores.

85%of the airnb cost in between 0-250 cad per night.

We see the distribution for pricing is strongly skewed right. This again makes sense as a majority of the listings on Airbnb are single individual listings. Additionally, Airbnb does strongly cater to travelers who are looking for cheaper places to stay for short durations of time. There are of course listings with a high pricing as well; intuitively this matches with hosts that are listing a high value property such as an entire house. To compensate for the skewed right distribution, we will log the response variable and store the results in a new panda column. We will then run our baseline models on both the logged and original response and compare the results.



Airbnb Price Prediction:

According to radial column chart the accuracy of our 5 models are:

1. Airbnb clustering
2. Airbnb price prediction
3. Calender data analysis
4. Words of the seller

##Calender CSV Analysis:

We are going to look at Airbnb listings and calendars, and trying to provide some exploratory analysis around predicting listing prices, both for, if we were hypothetically working at Airbnb, and also for a consumer. Let’s get started! We have 1619 days and 34255 unique listings in the calendar data and Timestamp starts expands between '2015-06-07’ and '2020-01-12’. Our reorganized calendar data covers six-year time frames with price and availability every day. **So, first we want to see what is the most popular times of the year for rentals of Toronto airbnb.**

For a Toronto homeowner considering renting out a home on AirBnB, it would be useful to know the most popular times of the year for rentals of a Seattle home. This would enable them to plan the timing for any preparation of the home for the peak season, as well as planning maintenance or upkeep work for less popular months.

**Availability on the Calendar**

When we look at calendar data, we may want to ask questions like: how busy will it be for Airbnb hosts in Toronto for the next year? Among the total listings over six years 13591002 were not available and 10106623 were available. If we look at it by year segmentation, we see that:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Start Date | End date | Available | Not Available | % Availability |
| 2015-06-07 | 2016-06-05 | 1487547 | 622883 | 70.5% |
| 2016-07-05 | 2017-07-04 | 1973857 | 1751698 | 52.9% |
| 2017-06-03 | 2018-06-02 | 1988684 | 2651926 | 42.8% |
| 2018-07-06 | 2019-07-06 | 2007029 | 4040656 | 33.2% |
| 2019-01-13 | 2020-01-12 | 1487547 | 622883 | 70%(Predicted) |
| **2015-06-07** | **2020-01-12** | **10106623** | **13591002** | **42.6%** |

To find out daily average availability for one year, we have converted available column formatting to 0 if available and 1 if not busy. The plot is below:

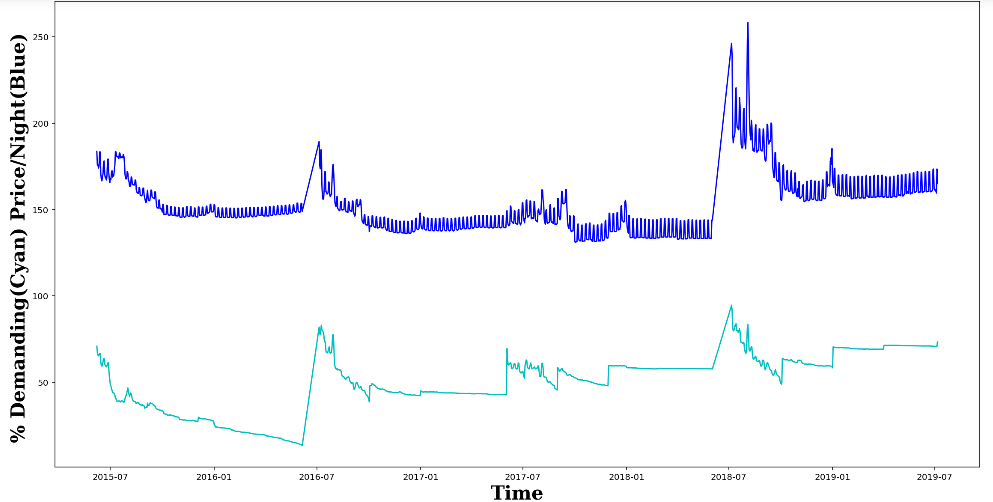


Figure :Clear spikes of number of non-available listings and prices from 2015 to 2019

Figure Findings:

Availability is not stable all year round.

Increase from Sep to Nov.

Remain relatively stable from 01/2017-09/2017, although the reason why there are two sudden drop of home supply is unclear.

Average prices:

Price drop as home supply increase from 09/2016 to 12/2016.

The sudden drop of supply in 03/2017 dose not drive price up; However, the sudden drop of supply before 05/2017 just rocket price up. So I guess it has something to do with demand change.

There seems to have a small periodical price moving circle and this maybe correspond to weekends. It shows the percentage of listing listings that become unavailable in Toronto per day, through 2015 to 2019 with certain spikes at certain times of the month. Moreover, the highest unavailability is of listings occurs near in July to October. The busiest month in Toronto is always the October. The next busy months seems after April and extend to the summer. These are all within our experience and expectations.

**Price on the Calendar Or seasonality analysis**

Next, we try to find out How price changes over the year by month. We remove “$” symbol in price column and convert it to numeric, and convert date to datetime data type.

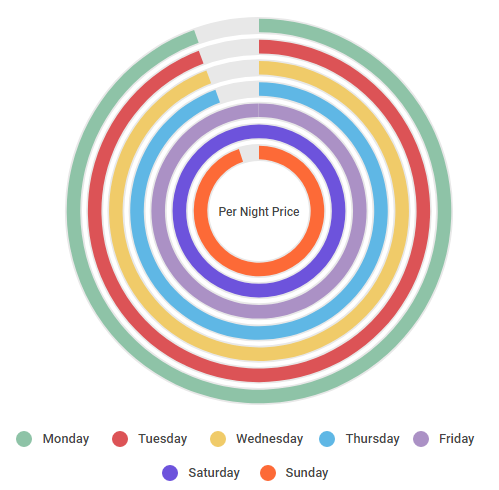
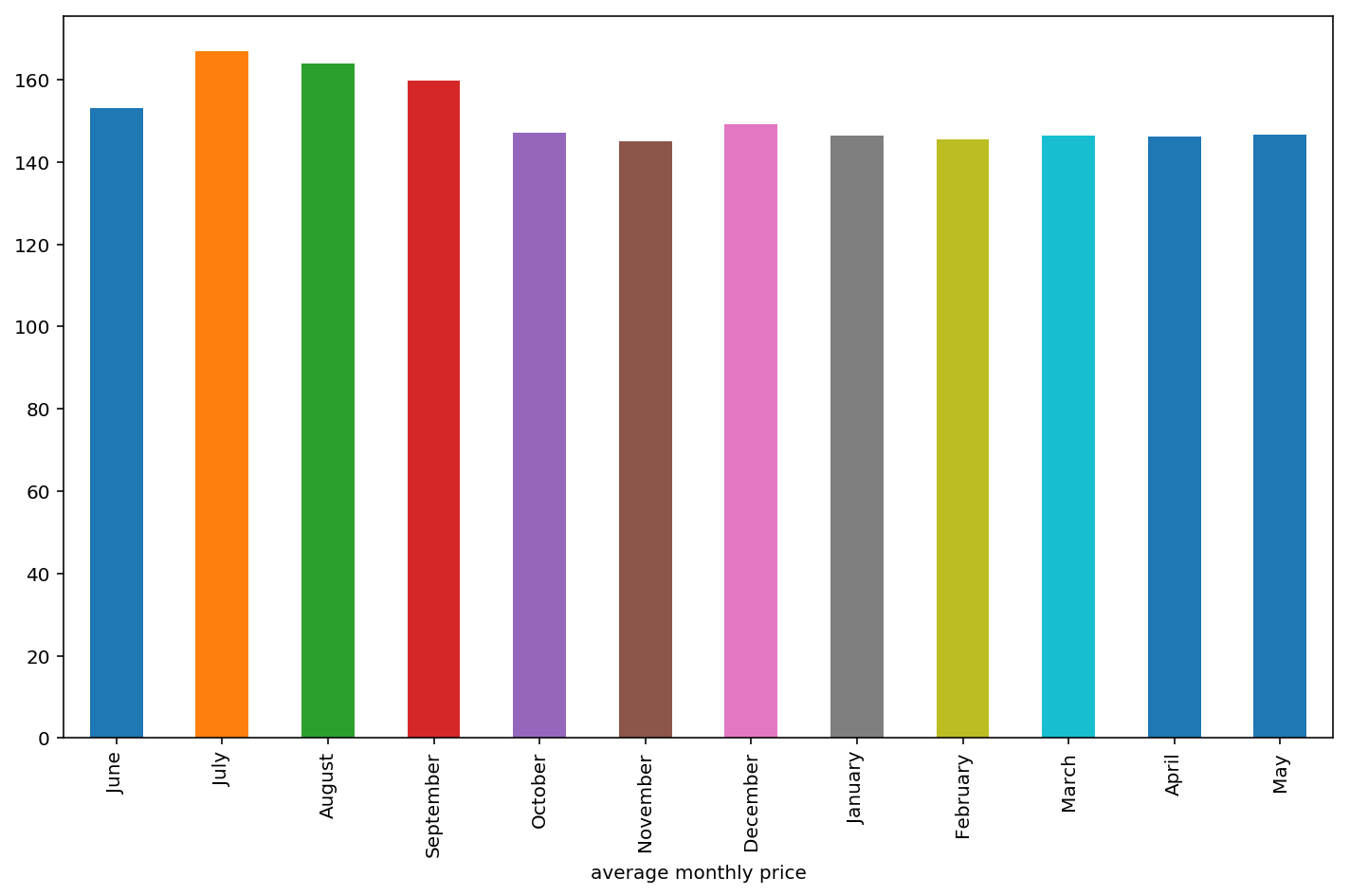


Figure : Weekdays variation of Airbnb Per Night Price



The Airbnb price in Toronto increases in the months of July, August, September, October and December. Agreed, these three months are the best months visiting Toronto.

Business Growth Trend:

**Part III: When and where are the highest revenue-generating times of the year for Seattle homeowners?**

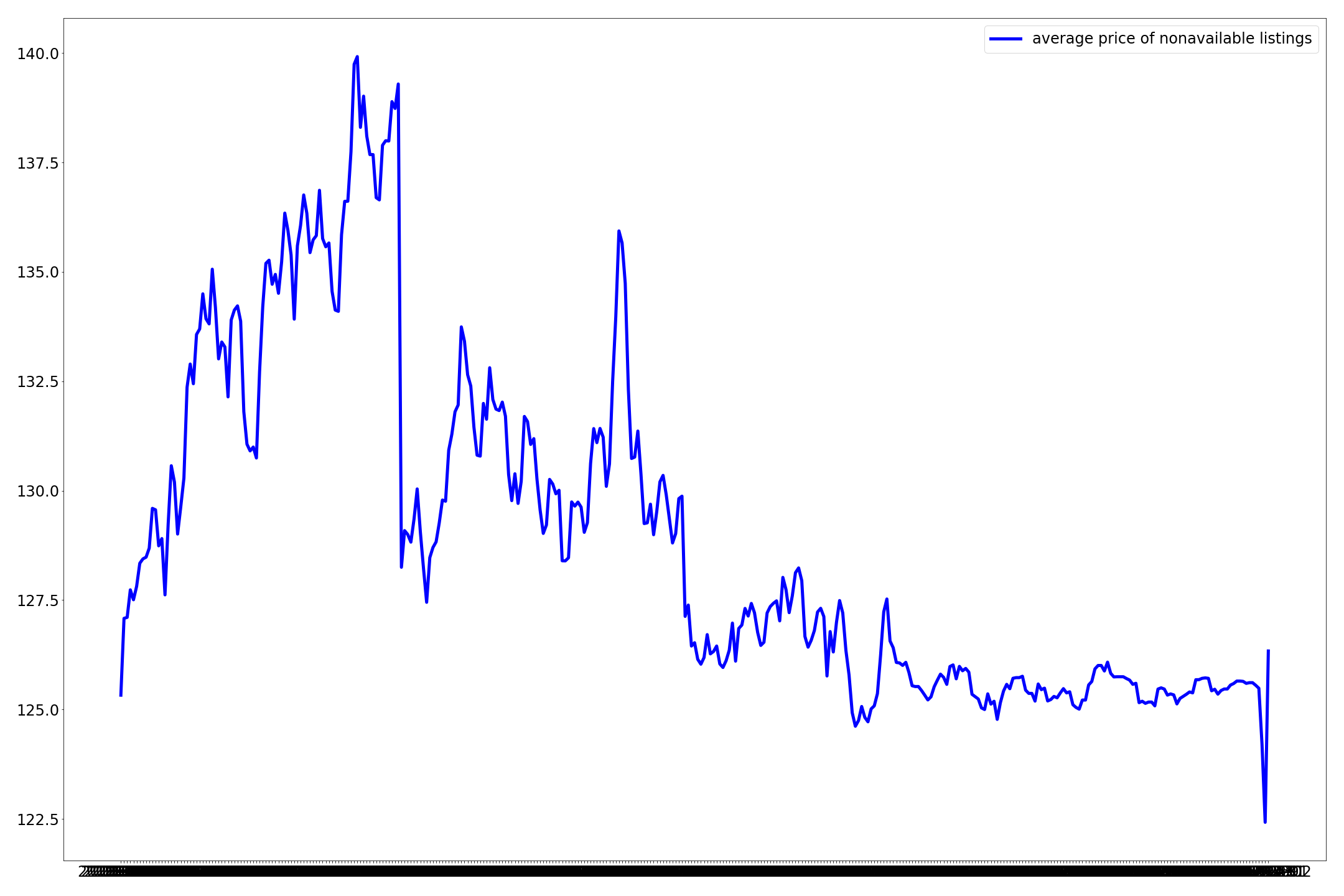
As a Seattle homeowner, my motivation for renting out my home is to earn revenue from rentals. Hence, it is worth looking into the data to identify the highest revenue-generating times of the year. If the data could show me what the average prices are for the peak times, then I would be able to set prices competitively and maximize my revenues from renting out my property then.

When analyzing the data, I assumed that if a property is rented at the price of that date (hence it is unavailable), it means that that price is the market price, as it is the price that is willingly paid for by a guest. Conversely, if a listing is available during that period, this would imply that the price is not willingly accepted by the market.

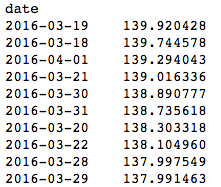
This assumption follows the real estate definition of a home’s market price, as the price agreed upon by a willing renter and a willing homeowner (from[here](https://budgeting.thenest.com/market-value-vs-market-price-4397.html)).

These may be simplified assumptions, but I decided that given the motivation of the analysis, they are acceptable simplifications. As a Seattle homeowner, I am more interested in the days of the year that guests/renters are willing to pay higher rent for my home, so that I can list a competitive price during that time. There is less of a priority to know all the other factors that may lead to a price being higher in that particular time.

The chart below charts the average price of all unavailable listings per day from 2016 to first two days of 2017. We can see that the **peak average prices occurred around March 2016**, while prices have started to drop towards 2017.



The average price for non-available listings has the highest spike in March 2016, and has been slowly decreasing towards the end of the year

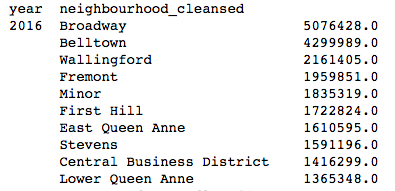


Average prices for non-available listings range from $138–$140 in March 2016

Pulling out the specific average prices in the month of March in 2016, we can see that the price ranged from **$138–$140**. However, as we have no data for March 2017, we are unable to check if a similar price range holds in the next year.

We would need more data to validate if March is a seasonal peak for home rentals on AirBnB in Seattle, but this is a good start for a Seattle homeowner to explore the peak and trough seasons for home rentals.

Meanwhile, to identify the areas which would generate the most revenue for a Seattle homeowner, I combined the analysis from Part II to Part III. It is not necessarily the case that the highest priced area would give the most revenue to a homeowner, as the location may not be popular with guests. Hence, combining it with Part II analysis where I can identify where are the most popular areas, I can identify the highest generating revenue areas.

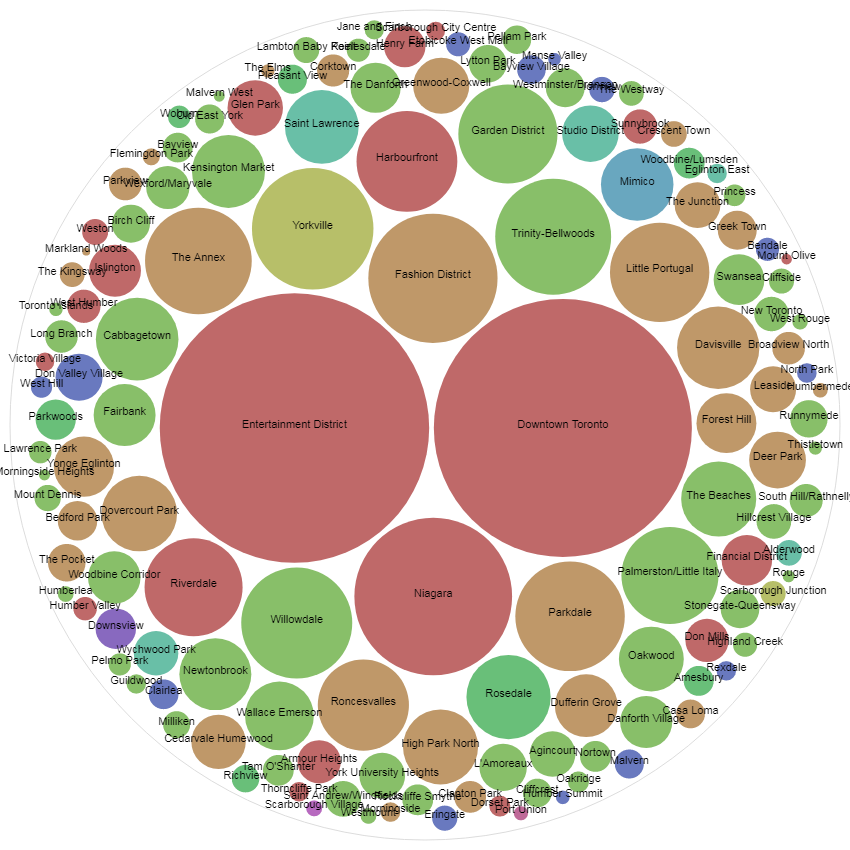


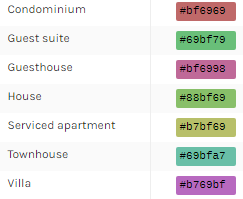
In the year 2016, the highest amount of revenue based on the number of rentals x the price of a listing is in Broadway neighbourhood

After combining the two analyses, I found that in the year 2016, the **Broadway neighborhood** was the clear top revenue generating neighborhood for Seattle homeowners, while Belltown is a distant second. Meanwhile, the other neighborhoods are about 60% less revenue-generating compared to Broadway.

Hence, an actionable next step for the prospective AirBnB homeowner is to check which neighbourhood their home is in and see what is the competitive prices for their home in that time of year, while checking how popular their neighbourhood is.

That being said, an important caveat for this analysis is that the data is only for one year’s worth of house listings. The trends might be different for 2018.





Price per night according to area

**Neighbourhood Word Cloud:**

The sellers are focusing their neighbourhood zones for Airbnb posting. The word cloud regarding the neighbourhood gives us an interesting insight into the area. We dug through 15130 words with log n scale and displayed most cited 400 words.



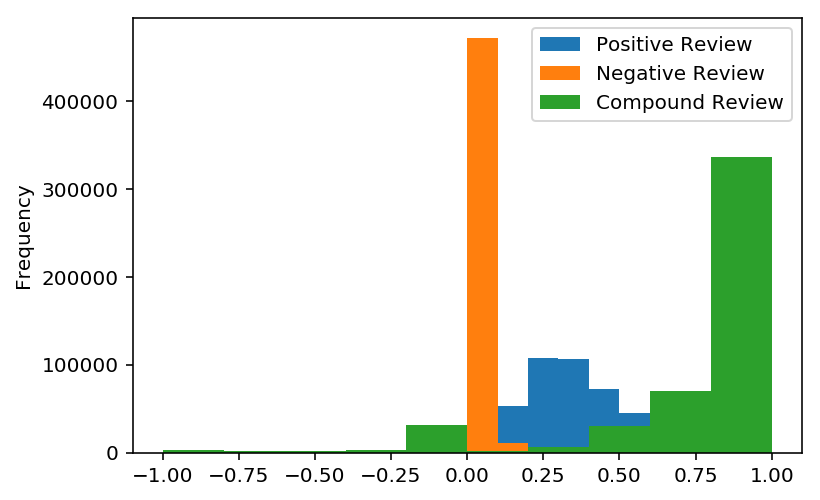
It can be seen that unique words like cozy, Toronto, Car, Walking, Restaurant, close, city, min are very commonly words used when hosts are describing their homes. This is making sense because the primary purpose of airbnb is not to provide luxury hotel suites but just a suitable place to stay and travel easily to the center of the city. Naturally hosts understand these purposes and create their summaries based on location and requirements in order to attract as many travellers as they can. So if hosts are not able to attract too many travellers then they can add these keywords to their summaries in order to attract travellers to choose their listings.

# **Analysis-4:**

# **Sentiment Analysis Of Reviews & its relationship with Price**

There are so many factors which contributes towards the price of a listing on AirBnB.While, we already have few conclusions for relationship between various factors and their dependency on prices of a listing,lets analyze if price of a listing dependent upon number of reviews and if yes, how does it varies?

# 



We used publicly available pre-trained Vader sentiment model based on NLTK to check what the reviewers’ sentiment about the listings. The overwhelming majority of reviews (>80%) according to density plot appears to be positive. Looking deeply in the negative and compound 95reviews we see the clear multi-peaked structure. It can be due to the reason that the model simply misclassified them. 15% of the reviews are strongly negative, and 50% can be termed as a mixture of positive and negative.

Next, we want to make recommendations for the low earner hosts and make their listings more profitable and gain more revenue than before. This notebook do features comparison between listings that have high earning versus low earning listings (in term of # of reviews and rating of reviews). The features will be chosen based on their several top-rank correlation of a new feature called new\_score\_review. \*\*New\_score\_review\*\* is a feature of the multiplication of \*\*reviews\_per\_month\*\* and \*\*review\_scores\_rating\*\* and \*\*divided by 100\*\* (for the sake of simplicity). The creation of the new feature is intended as a approach to know which listings have more completed orders than someone else and great reviews from Guests as well since we dont have any number of completed order. Therefore, we could know what makes the top performer become top performer, and how the low performer one could adapt the behavior

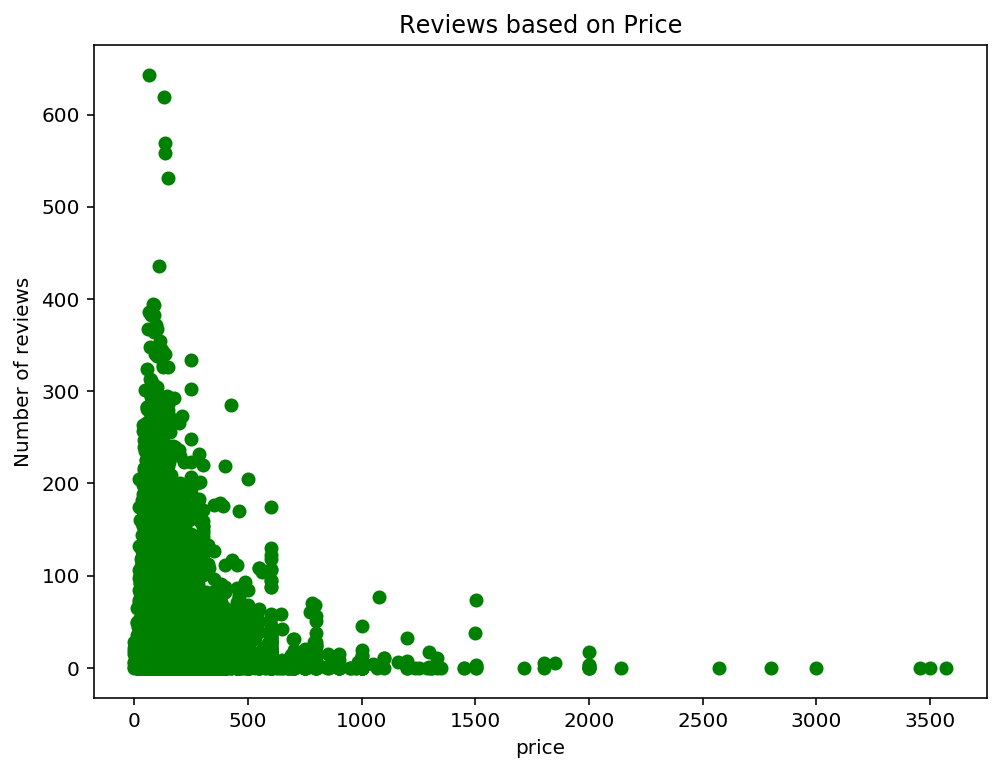
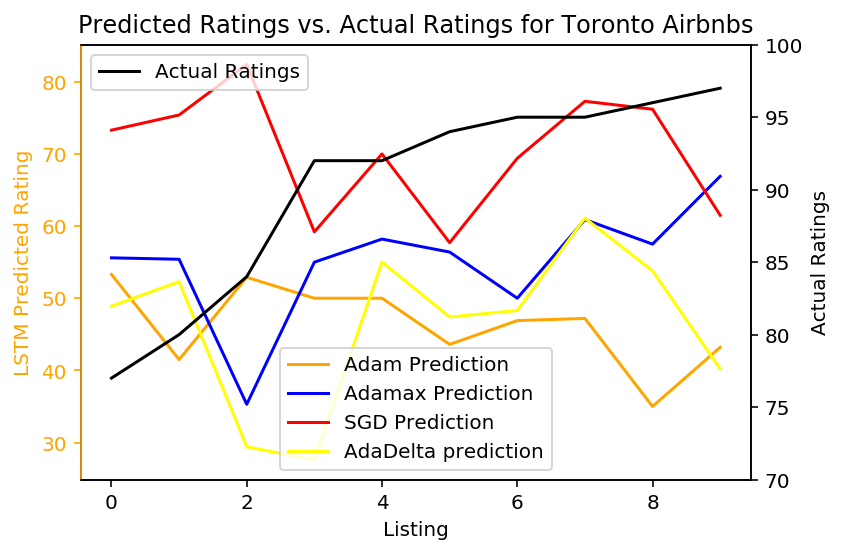


Figure : Review Depends On Price

The graph shows that listings with prices that range around 0 - 500 get the most reviews, probably because they are in the most reasonable price range. The number quickly declines as the price goes up. This indicates that more people book listings that are around $0 - 500 in prices. This shows that it is not necessary for an expensive listing to have large number of reviews. Heneforth there is no exact relation between Prices and Number of Reviews for a listing.

Number of unique reviews for particular id is a pretty big range from 1 to 677. As our data is very large, we are considering listings that have at least 100 reviews for prediction with LSTM neural network model. The LSTM model forecasts reviewer’s sentiment against the `listing\_id` and compare the actual rating. Here is a comparison of different optimizers of LSTM model, we see that SGD applies best to our case.



It can be seen that the most talked about words are "great location", "great host", "walking distance" and "highly recommended".All these reviews and comments plays a big role in attracting the attention of travellers and if there are comments such as "highly recommended" then travellers surely takes a look at the listing.

Top Earner Vs Low Earner:

\*\*Insights\*\* of features comparison (from top to bottom):

\* Acceptance Rate:

- TP: Almost all of the top listings always accepts bookings.

- LP: Almost half of them never really accept the bookings.

\* Host Identity Verified:

- TP: Almost 90% of the host’s listings has been identified by Airbnb.

- LP: 38% of host’s listings hasn't been identified by Airbnb.

\* Superhost:

- TP: 40% Top listings have Superhost Predicate.

- LP: Most of the low performers are not superhosts.

\* Instant Booking Feature:

- TP: Eventhough, have big contribution on the score most of the top listings don’t activate their instant\_booking by ~62%

- LP: More than 90% they turn off the instant bookable.

\* Respond Time:

- TP: ~78% of top listings’ hosts always response conversations immediately.

- LP: Many of host of low performer listings respond the conversation longer than top-performer listings.

\* Respond the Conversations:

- TP: Most of the top listings in Seattle always response conversations.

- LP: A lot of low performer never response the conversation.

\* Amenities:

- TP & LP: They serve assorted miscellaneous in the top listings; most of them provide wireless internet, heating, essentials, smoke detector, shampoo, etc. There are not much difference between low performer and top performer listings about the amenities.

\*TP: topperformer | LP: lowperformer\*

# Suggestions for Hosts.

\*Therefore,\* to conclude the analysis there are several points that low-performer could do to increase their probability to get more bookings, reviews, and higher number of reviews;

1. \*\*Increase the acceptance rate of the rental.\*\* Top performer almost never reject the orders, not like low performers around 45% of the listings that did not accept any bookings. Although we do not know the reason why the hosts did not accept the orders;

2. \*\*Be responsive\*\* most of the top performer hosts always giving response within an hour about 80% of all the time.

3. \*\*Be a superhost.\*\* It’s the status and the recognition from the Airbnb because they provide amazing experience and great example for other hosts.

4. \*\*Always response the conversations.\*\* More responsive the hosts, more better the score it would be. Top performers always response to every conversation.

5. \*\*Activate the instant bookable features;\*\* They give better experience to the future guests.

6. \*\*Make your account verified by Airbnb.\*\* Many of low performer accounts have not verified by Airbnb more than the high performer by 25%.

\*Future development plan: \*This new comprehensive recommendations feature, could be attached in every page of User Interface that have Informations about those variables, notifications, or other forms of reminders to the eligible hosts. Hosts that subscribe this feature could be charge in reasonable price.!

<https://www.kaggle.com/yogi045/how-to-become-top-earner-in-airbnb>

<https://www.kaggle.com/ibjohnsson/predicting-listing-prices>

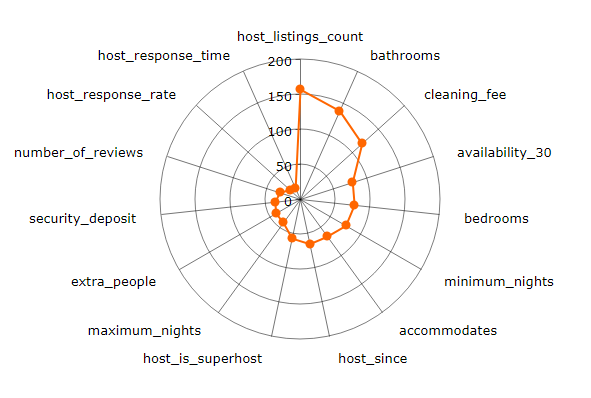


Figure : Scores of Important Features For Price Prediction According to XGB Model

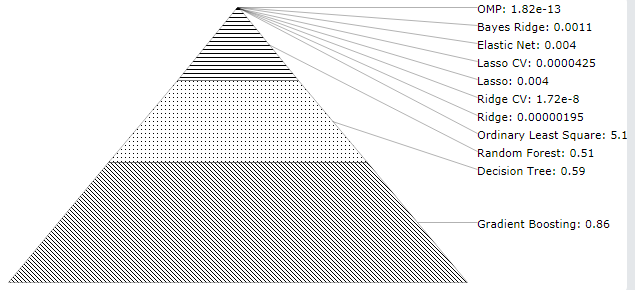


Figure : Pyramid Chart showing the MAE value of different regression model for predicting price

As we see, our MAE values are very small, specifying our perfect model accuracy. The topmost models perform the best.

https://www.kaggle.com/aleksandradeis/airbnb-seattle-reservation-prices-analysis

https://www.dataquest.io/blog/understanding-regression-error-metrics

# **ANALYSIS - 5**

# **HOST ANALYSIS & RECOMMENDATION SYSTEM FOR PRICES**

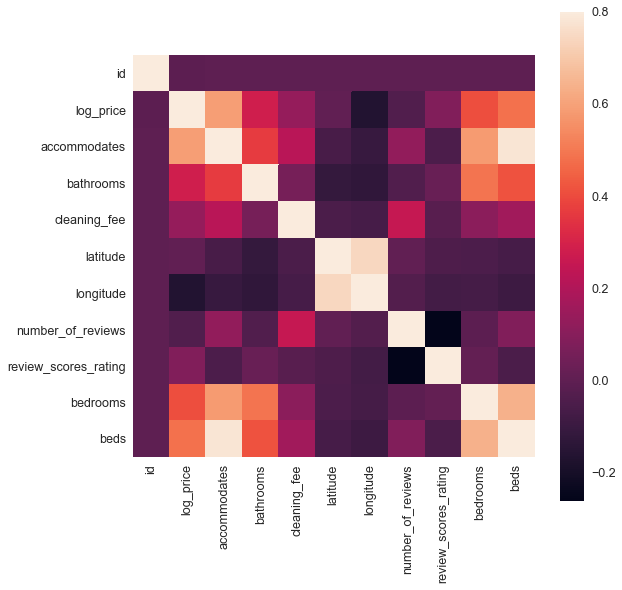
Lets analyze the trend of people hosting AirBnBs in Toronto. We can get insights about how much more popular is AirBnB among hosts now than it was two or three years ago.

Prediction:

1. Price Prediction
2. Review Prediction
3. Neighborhood Prediction

References:

1. <https://www.jasondavies.com/wordcloud/>
2. <https://bl.ocks.org/kerryrodden/766f8f6d31f645c39f488a0befa1e3c8>
3. <https://towardsdatascience.com/exploring-machine-learning-for-airbnb-listings-in-toronto-efdbdeba2644>
4. <https://bl.ocks.org/kerryrodden/7090426>
5. <https://datavizcatalogue.com/methods/radial_bar_chart.html>
6. <https://github.com/ruchigupta19/Boston-Airbnb-data-analysis/tree/master/Analysis>
7. https://live.amcharts.com/new/edit/

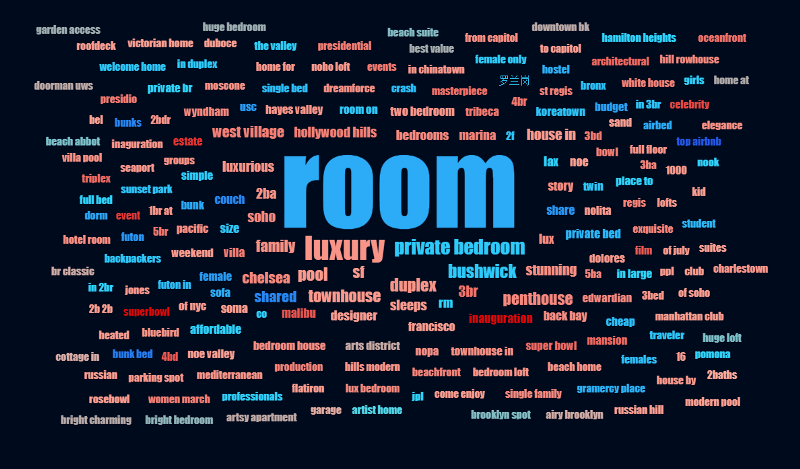


Correlation matrices provide a quick and easy visual to make sense of multiple variable interactions.

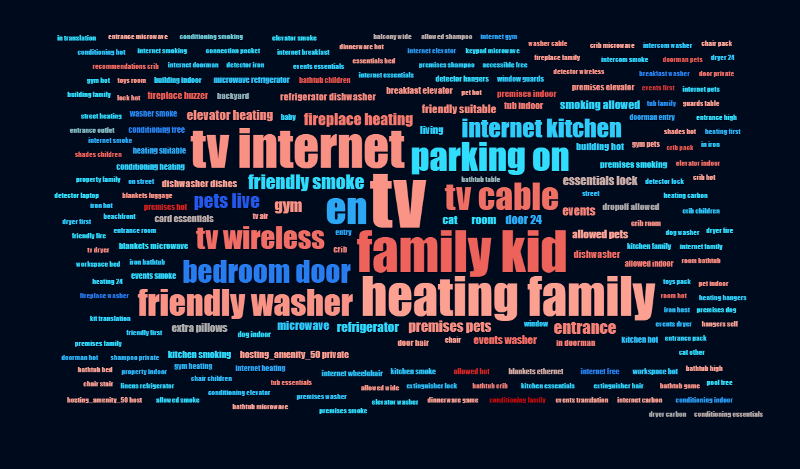
In this figure, the lighter colors represent high correlations with the intersecting features. Our attention is drawn to the accommodates, bathrooms, and cleaning\_fee as features that are highly correlated to log\_price. This makes sense based on our initial observation, although the high importance of cleaning\_fee is not something we predicted. Another benefit is we are able to see features that are highly correlated with one another, and thereby may impact our model creation. Latitude & longitude and beds & bedrooms are highly correlated as expected. Let’s note this and come back to this observation later when building our model.

These correlation matrices are good for numerical values but do not give much insight as to what’s happening in our categorical and text fields. Generally, I will replace categorical features with dummy variables (NYC = 1, Boston = 2, etc). For unstructured text, we must take a different approach. Qualitatively, it would be nice to get a sense of the typical content of these features. Luckily, we can make a word cloud that will help us out — let’s take a look:

**Name**



**Amenities**



**Description**



We can see how word clouds allow users to quickly identify the most frequent tokens that appear in each feature. By breaking down each string into separate components, we are able to identify common words that appear over and over again. Unsurprisingly, “room”, “bed”, and similar qualifiers are the most frequently used across categories. I’ll leave it as an exercise to determine how this observation can be successfully incorporated (consider frequency).

‘name’, ‘summary’, ‘space’, ‘description’, ‘neighborhood\_overview’, ’transit’, access 12094 non-null object

house\_rules 12882 non-null object

host\_id 19255 non-null int64

host\_name 19252 non-null object

host\_since 19252 non-null float64

host\_location 19240 non-null object

host\_response\_time 19252 non-null object

host\_response\_rate 19252 non-null float64

host\_is\_superhost 19252 non-null object

host\_neighbourhood 16674 non-null object

host\_listings\_count 19252 non-null float64

host\_total\_listings\_count 19252 non-null float64

host\_verifications 19255 non-null object

host\_has\_profile\_pic 19252 non-null object

host\_identity\_verified 19252 non-null object

street 19255 non-null object

neighbourhood 18161 non-null object

neighbourhood\_cleansed 19255 non-null object

city 19254 non-null object

state 19244 non-null object

zipcode 18836 non-null object

market 19213 non-null object

smart\_location 19255 non-null object

country\_code 19255 non-null object

country 19255 non-null object

latitude 19255 non-null float64

longitude 19255 non-null float64

is\_location\_exact 19255 non-null object

property\_type 19255 non-null object

room\_type 19255 non-null object

accommodates 19255 non-null int64

bathrooms 19242 non-null float64

bedrooms 19248 non-null float64

beds 19236 non-null float64

bed\_type 19255 non-null object

amenities 19255 non-null object

price 19255 non-null int64

cleaning\_fee 15877 non-null float64

guests\_included 19255 non-null int64

extra\_people 19255 non-null int64

minimum\_nights 19255 non-null int64

maximum\_nights 19255 non-null int64

calendar\_updated 19255 non-null object

has\_availability 19255 non-null object

availability\_30 19255 non-null int64

availability\_60 19255 non-null int64

availability\_90 19255 non-null int64

availability\_365 19255 non-null int64

calendar\_last\_scraped 19255 non-null int64

number\_of\_reviews 19255 non-null int64

first\_review 15526 non-null float64

last\_review 15527 non-null float64

review\_scores\_rating 15262 non-null float64

review\_scores\_accuracy 15250 non-null float64

review\_scores\_cleanliness 15252 non-null float64

review\_scores\_checkin 15247 non-null float64

review\_scores\_communication 15253 non-null float64

review\_scores\_location 15244 non-null float64

review\_scores\_value 15246 non-null float64

requires\_license 19255 non-null object

instant\_bookable 19255 non-null object

is\_business\_travel\_ready 19255 non-null object

cancellation\_policy 19255 non-null object

require\_guest\_profile\_picture 19255 non-null object

require\_guest\_phone\_verification 19255 non-null object

calculated\_host\_listings\_count 19255 non-null int64

reviews\_per\_month 15526 non-null float64