**Airbnb Review Sentiment Analysis and Price Prediction**

George Janjalia

Baskar Dakshinamoorthy

**Introduction:**

Airbnb has seen a meteoric growth since its inception in 2008 with the number of rentals listed on its website growing exponentially each year. Airbnb has successfully disrupted the traditional hospitality industry as more and more travellers, not just the ones who are looking for a bang for their buck but also business travellers resort to Airbnb as their premier accommodation provider.

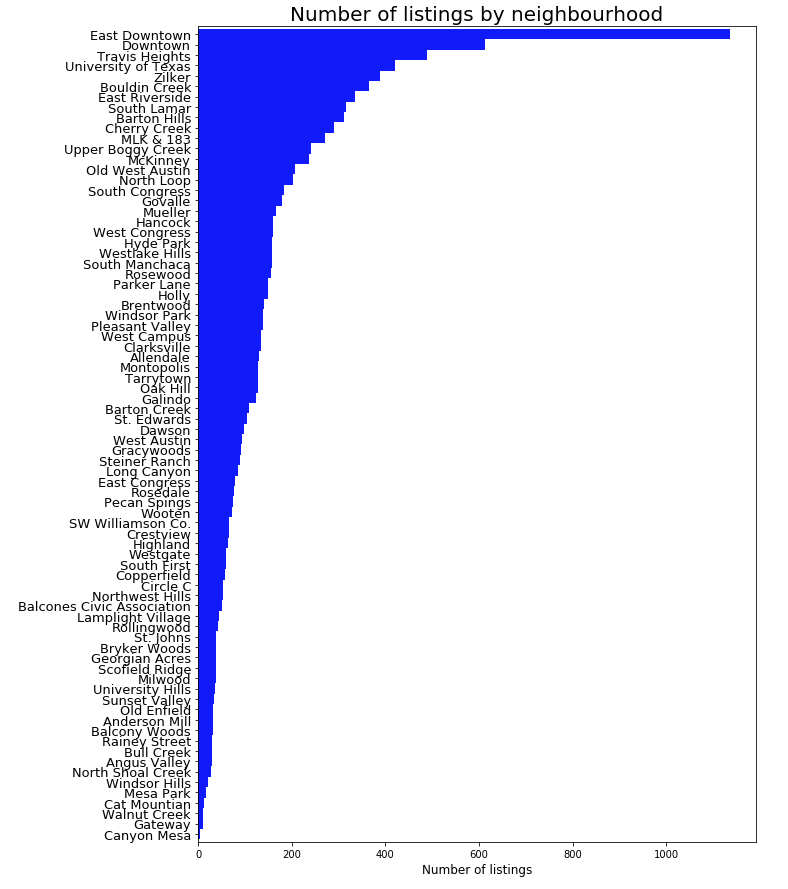
**Dataset**

[Airbnb](https://www.airbnb.ca/?locale=en) does not provide open data in the sense of giant databases or dumps that we can work with. However, [Inside Airbnb](http://insideairbnb.com/index.html) utilizes public information compiled from the Airbnb web-site and analyzes publicly available information about a city’s Airbnb’s listings, and provides filters and key metrics so we can see how Airbnb is being used in the major cities around the world. [Inside Airbnb](http://insideairbnb.com/index.html) is an independent, non-commercial set of tools and data that is not associated with or endorsed by Airbnb or any of Airbnb’s competitors.

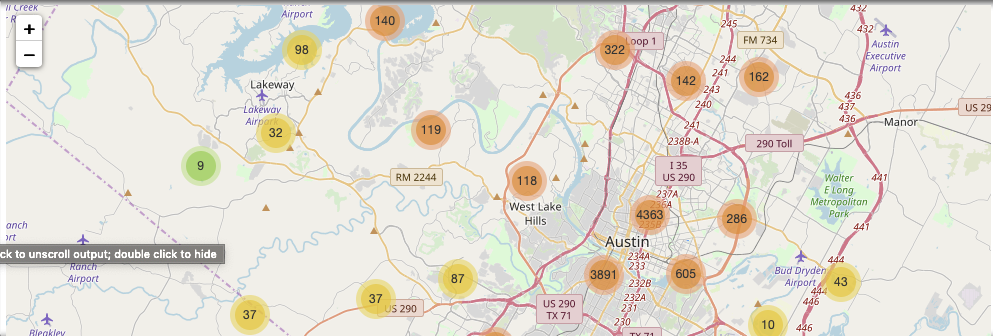
We grabbed the Austin listings and reviews data from Inside Airbnb as .csv files and loaded them into Pandas DataFrames, then joined them on listing\_id. The joined DataFrame provided 103 fields to work with, including the date of the review, id of the listing, neighborhood the listing is in, text of the review, and many other pieces of information. Then, I wrote and applied a series of scripts to clean and process the raw text data into a format I could more easily work with, including tokenizing reviews at the sentence and word levels and grouping reviews by neighborhood.

**2. Data exploration:**

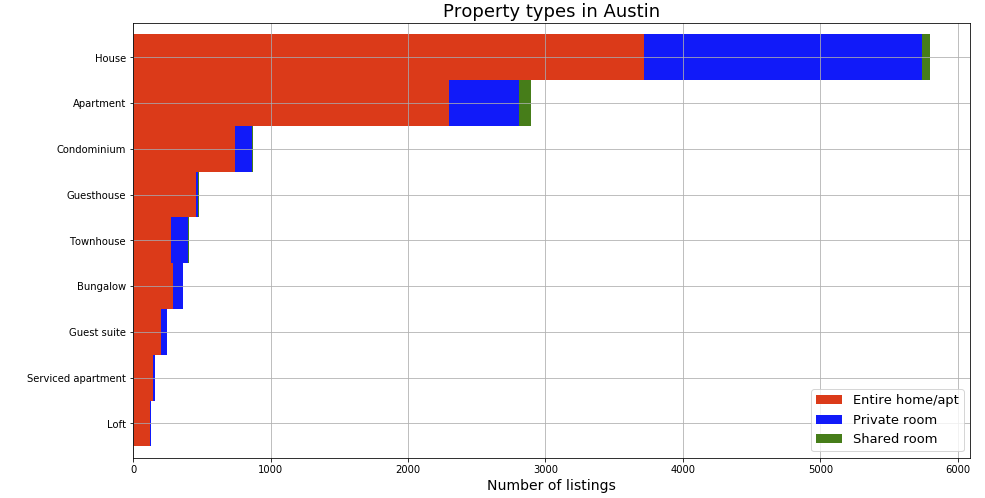
Based on the listings East Downtown and Downtown has more than 1000 listings which is understandable as tourists tend to stay close to the city attractions.



Below is the In-Map representation of all listings plotted in map. You can see that most listings are in the city center .

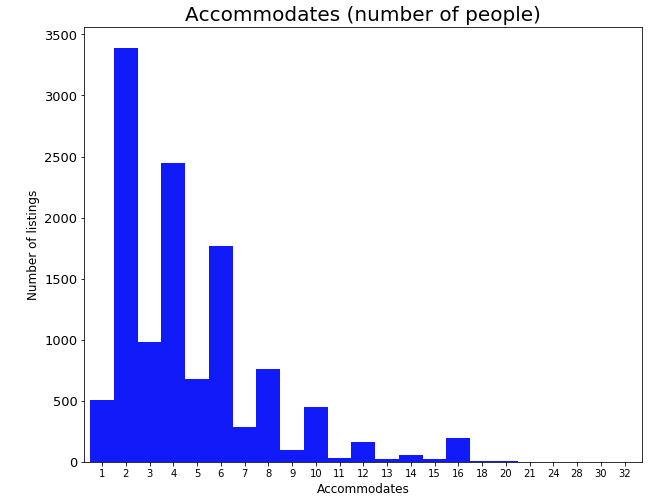


**2.2 Room types and property types**

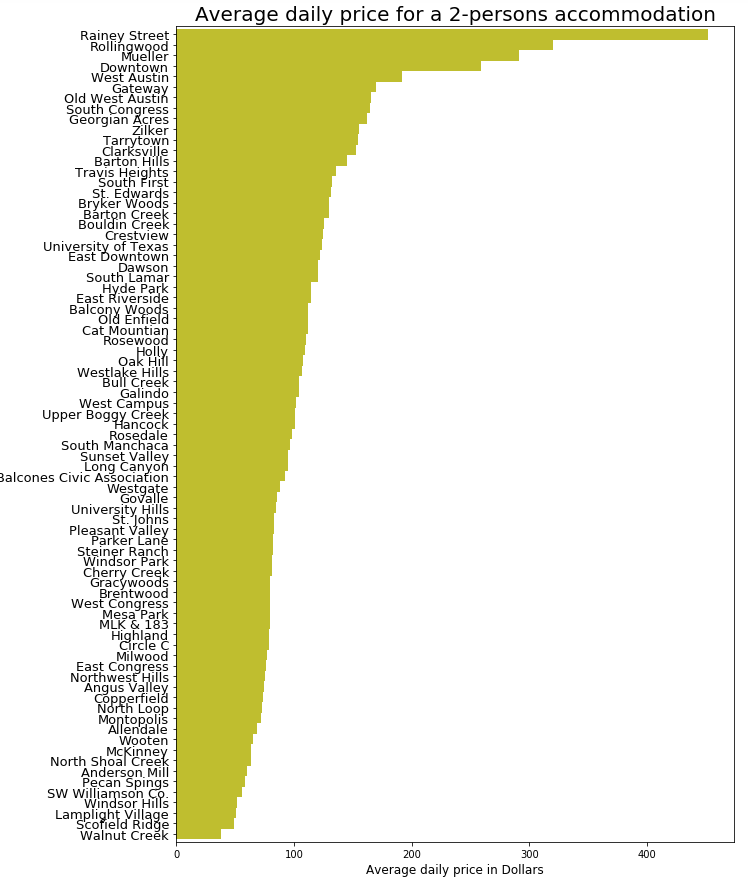


**Accommodates (number of people)**

As expected, most listings are for 2 people. In addition, Airbnb uses a maximum of 16 guests per listing.



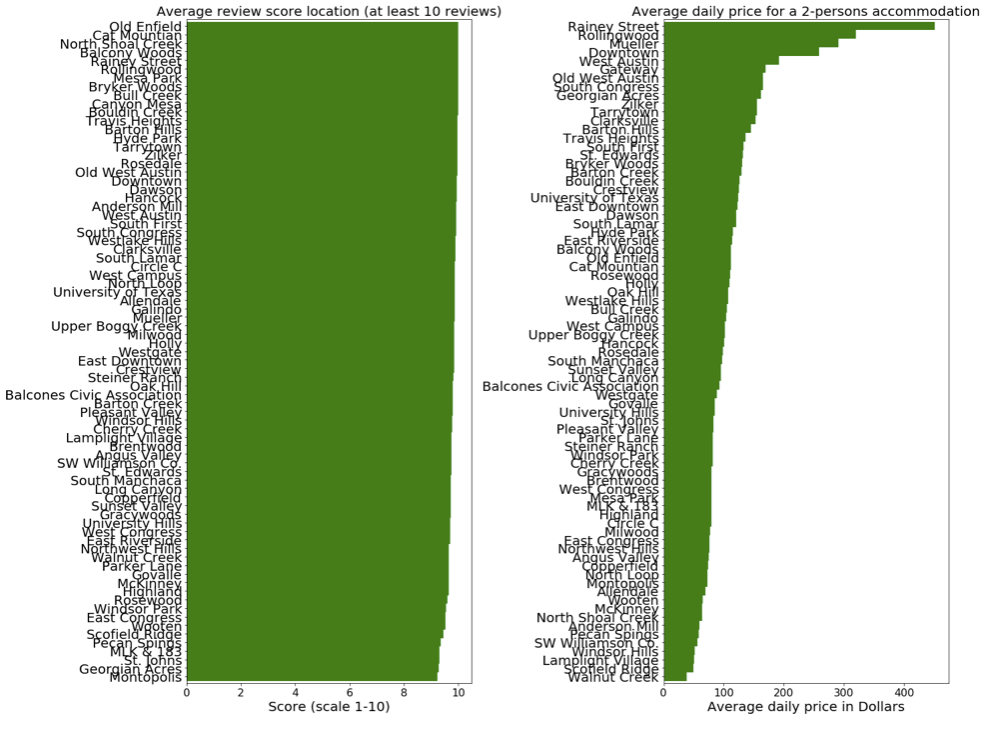
Average daily prices for 2 person accommodations seems to be very high on Rainy Street,Rollingwood and Muelller



**Review scores location, and location scores versus price:**

In this section, I am grouping the review scores for the location by neighbourhood (only listings with at least 10 reviews). Although I expect the distance to the city centre to an important factor, these score should also take other things into account. Other factors may include:

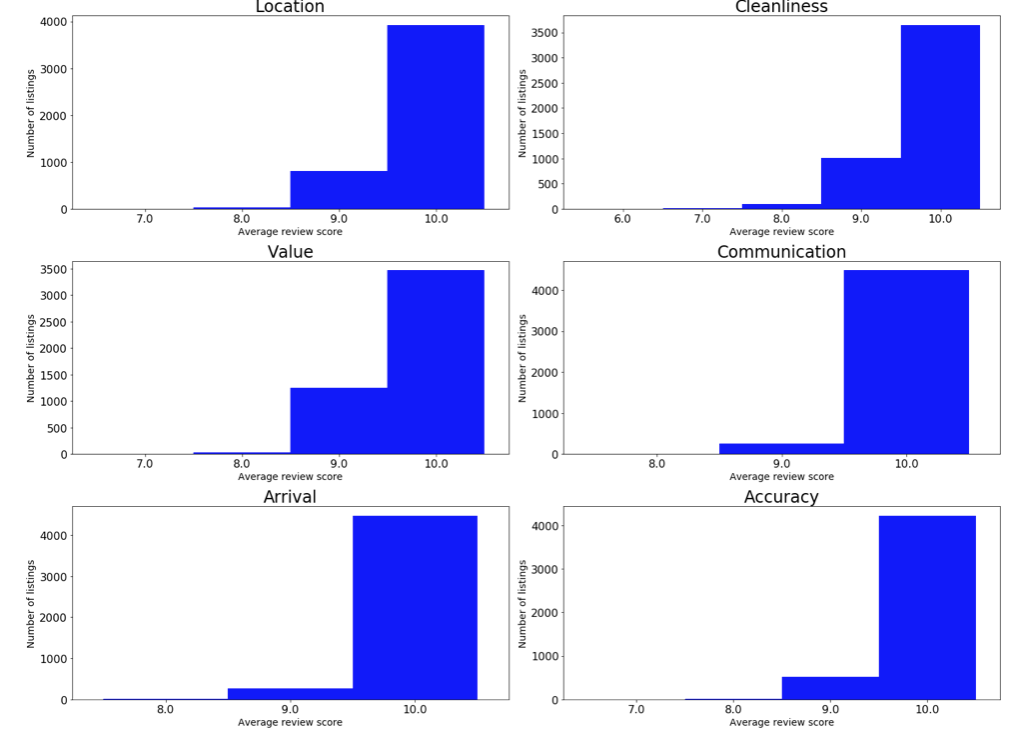
The safety of a location. Noise,If a listing is centrally located, but surrounded by noisy bars, that should cost point in the location review score. If a listing is located outside the city centre but well connected by public transportation, it should get bonus points for that. Facilities near the listing. Are there any supermarkets, bars and restaurants nearby? Some people may be looking for free parking, if they come by car.



**How to use review scores:**

In addition to written reviews, guests can submit an overall star rating and a set of category star ratings. Guests can give ratings on:

Overall Experience. What was your guest’s overall experience? Cleanliness. Did your guests feel that your space was clean and tidy? Accuracy. How accurately did your listing page represent your space? Value. Did your guest feel your listing provided good value for the price? Communication. How well did you communicate with your guest before and during their stay? Arrival. How smoothly did their check-in go? Location. How did guests feel about your neighborhood? Below you can see the scores distribution of all those categories. What caught my eye immediately is that scores seem really high across the board! A quick internet search told me that this seems common across Airbnb.

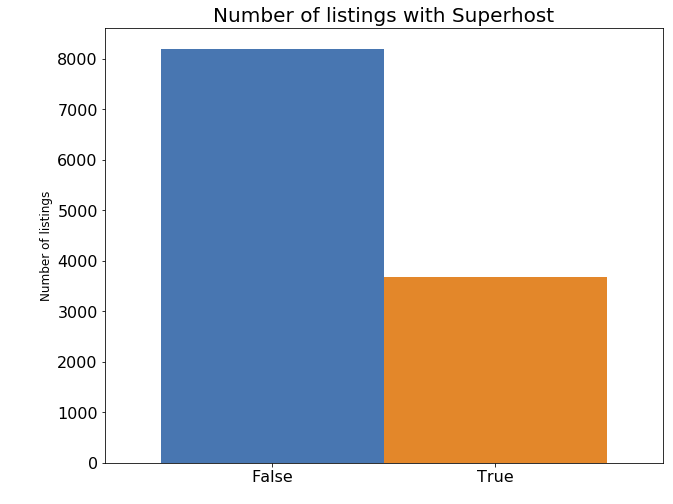


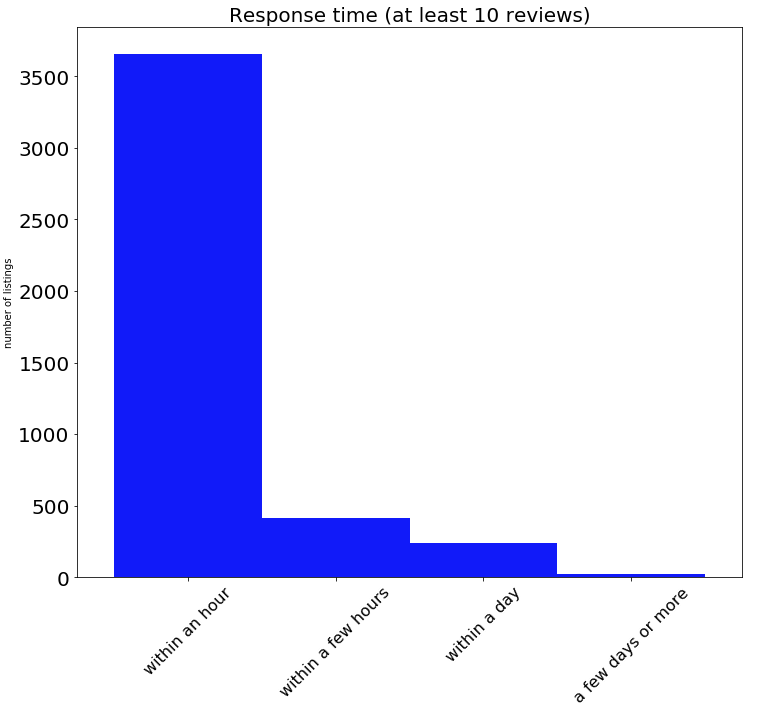
**Number of Listings with Superhosts:**

At Airbnb you can get the status "Superhost".

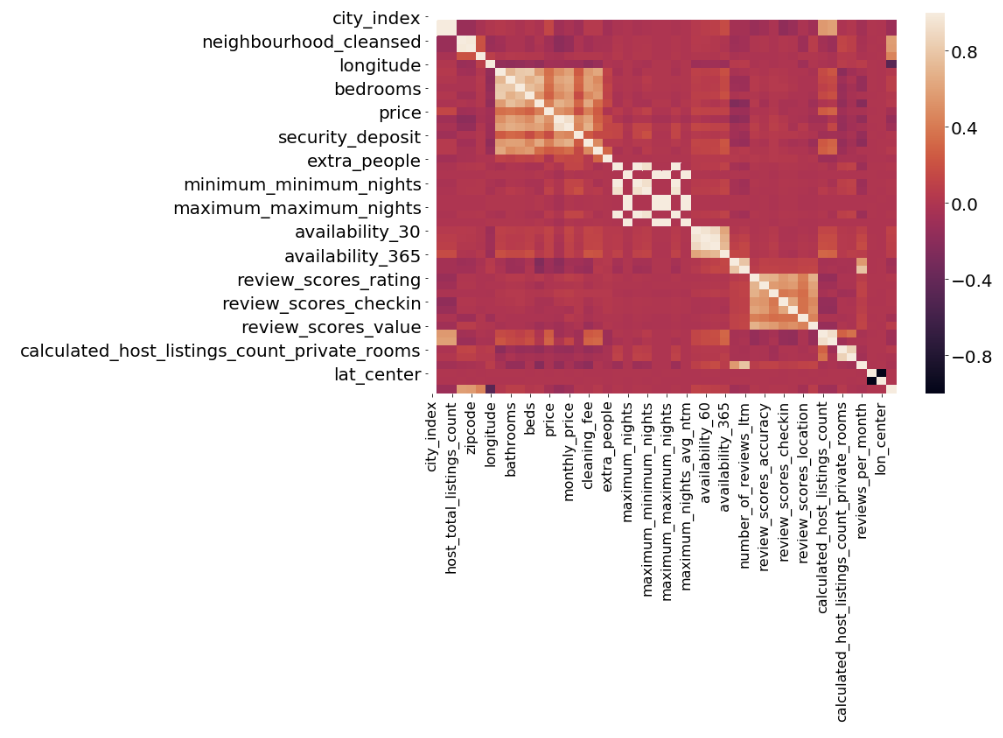
From Airbnb:

As a Superhost, you’ll have more visibility, earning potential, and exclusive rewards. It's our way of saying thank you for your outstanding hospitality. How to become a Superhost: Every 3 months, we check if you meet the following criteria. If you do, you'll earn or keep your Superhost status. Superhosts have a 4.8 or higher average overall rating based on reviews from at least 50% of their Airbnb guests in the past year. Superhosts have hosted at least 10 stays in the past year or, if they host longer-term reservations, 100 nights over at least 3 stays. Superhosts have no cancellations in the past year, unless there were extenuating circumstances. Superhosts respond to 90% of new messages within 24 hours.



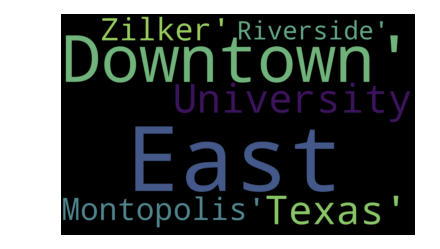


**Variable Correlation Heat Map:**



**Word Cloud on Reviews and No of Listings:**

Created wordclouds on Reviews and No of listings which shows most of the reviews seems positive in Austin and No of listings are more in East Downtown and Riverside.



**Sentiment Analysis:**

Now lets take a deep dive on Sentiment Analysis of Reviews and find the polarity of reviews.

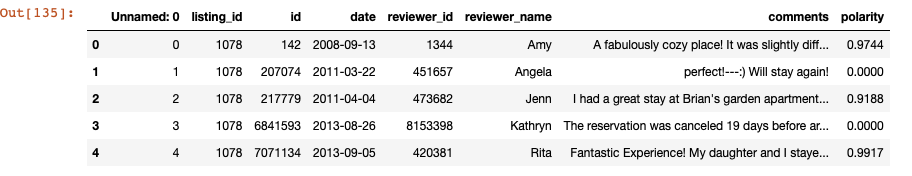
(Q1) what are the positive and negative reviews?

The field of our interest is comments that contains the textual comments. To transform it into review sentiments (to answer Q1), I will use the publicly available pre-trained [Vader sentiment model based on NLTK](http://www.nltk.org/_modules/nltk/sentiment/vader.html)

Compound sentiment polarity of the sentence came in the range of 0.8 to -0.4

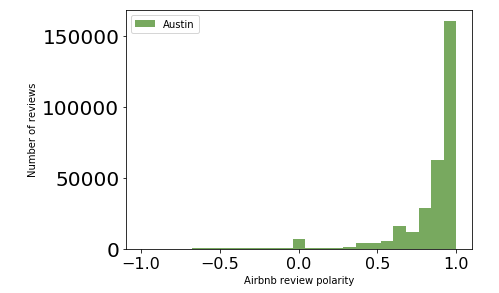
(0.8112, -0.4767)

Final sentiment polarity calculation for 845k entries takes about 2 hours

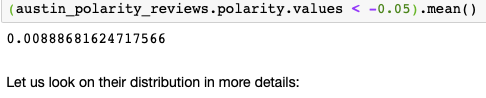


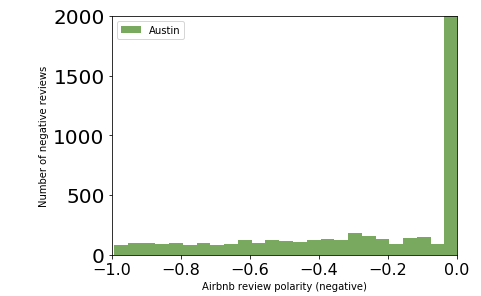
Plotting the histogram of review intensities reveals the following peculiarities:

* huge asymmetry between positive and negative reviews, 80+ k of them (or almost 50%) are close to maximally positive (polarity > 0.92);
* prominent peak of ~10k 'neutral' reviews.

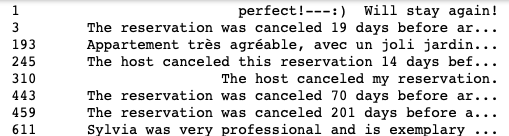


Importantly, very small fraction (< 5%) of all reviews of our interest are negative:





As we see, for datasets the top unique review polarity is 0.0. Lets Look closer for reviews with polarity 0.0:



Other comments made in English, such as 'The host canceled ...' or 'The reservation was canceled ...', are probably automatic messages generated by the system.

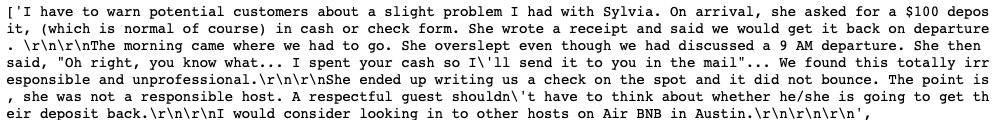
Created a function zero\_techinical\_reviews which Returns the fraction of "technical" reviews (staring from "The host canceled" or "The reservation was canceled") of zero polarity.



As we see, \*\*the technical messages fraction about apartment cancellation is very less.

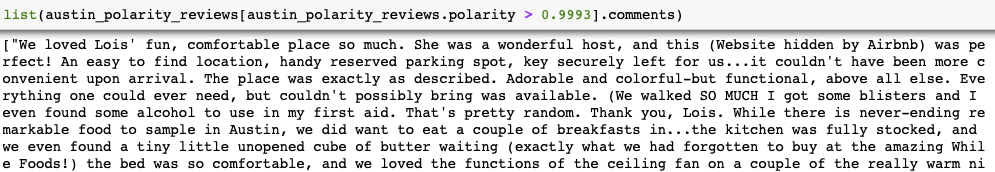
Now What about very negative reviews? First, look into the Austin dataset:

list(austin\_polarity\_reviews[austin\_polarity\_reviews.polarity < -0.8].comments.head())



Here, in addition to misclassified reviews in German like 'Vielen Dank an Jamie und Mike für die tolle Zeit in Austin', are some reviews in English with different (positive, negative or neutral) polarity that can also be interpreted as the model misinterpretation.

Finally, let us investigate the most positive reviews:



As we see, all top positive reviews are indeed very positive.

To summarize **Q1**, publicly available pre-trained [Vader sentiment model based on NLTK](http://www.nltk.org/_modules/nltk/sentiment/vader.html) is used to predict the sentiment polarity of recent reviews for Austin Airbnb listings . The model describes the sentiment polarity of English-written reviews fairly well but does not generalise for other common languages (German, French, Spanish etc.). As expected, the majority (95+% of all reviews) are positively- or neutrally-classified, and the large part of negatively- and neutrally-classified reviews are in fact misclassified positive reviews written in other languages. In addition to that, a subdominant part of neutrally-classified reviews are the typical technical messages due to host cancellation.

Models:

Three classifiers used are the Multinomial Naïve Bayes (MNB), Bernoulli Naïve Bayes (BNB) and linear Support Vector Classifier (SVC) models. To understand the BNB and MNB it is important to define Naïve Bayes first. The general term Naive Bayes alludes the solid independent assumption in the model, as opposed to the specific distribution of each component. A Naive Bayes model claims that every one of the attributes it uses are conditionally autonomous of each other given some class. Thus, on wants to calculate the likelihood of features 𝑓1 through 𝑓𝑛, given some class c, the Naive Bayes assumes the following:

This implies that when one needs to utilize a Naive Bayes model, the probability is a lot easier to work with:

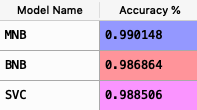
Obviously, these presumptions of independence are not true, yet practically speaking Naive Bayes models have performed very well, even on complex tasks where independence assumptions are false. Thus, the MNB is a particular occurrence of Naïve Bayes model, which for each uses a multinomial distribution (Difference between naive Bayes & multinomial naive Bayes, 2009).

The MNB is a supervised learning method, which requires labeled data and can be tested for accuracy using confusion matrices and cross validation (CV). The process of CV divides data in n number of folds (this case 10) and for each fold creates test and train sets. At the end of the process the model merges all n folds and provides average result (refer to figure 4 to view the example of cross validation). However, in order for the model to work data and labels are required to be fed in individually, thus as the model is created a data with no labels goes to one spot and labels go to another.

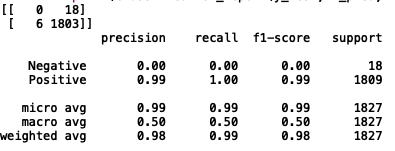
The BNB classifier is very similar to MNB, but instead of counting the accuracies of the words, it simply identifies presence of absence. Thus, if the word is present the dataframe will show 1 and if the word is absent the dataframe will show 0.

In data mining, support-vector machines are supervised learning models, which classify data and conduct regression analysis. Given a lot of training examples, which are assigned to any of the two categories, an SVM training model categorizes examples, making it a non-probabilistic double straight classifier. An SVM model is a representation of points mapped in space, which are separated by a reasonable gap that is as wide as could be expected under the circumstances. New models are then mapped into that same space and are predicted to fall on either one or the other side of gap.

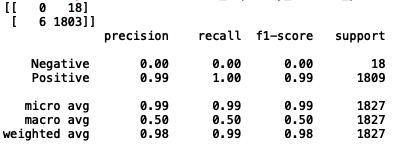
**Results:**

The results for all the models are pretty impressive – all of them are above 98%. With one glance it seems that the models are overfitted, however it is not the case. The reason, why the results are so impressive is that most of the reviews analyzed are positive, which leaves very little room for error. Also, not the full data was used. The authors randomly sampled 15% of the data because of the limited computational power. This means that not only there was very little amount of negative values, but possibly these negative values got filtered out leaving more positive reviews, as such as decreasing possibility of the error. Below are the full results for all three models:

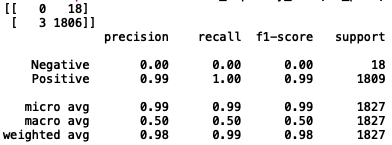
*MNB:*



*BNB:*

**

*SVC:*

**

**Conclusions:**

This study investigated the attributes that influence Airbnb users’ experiences by analyzing a “big data” set of online review comments through the process of text mining and sentiment analysis. Findings reveal that Airbnb users tend to evaluate their experience based on a frame of reference derived from past hotel stays. Three key attributes identified in the data include ‘location’, ‘amenities’ and ‘host’. Surprisingly, ‘price’ is not identified as a key influencer. Airbnb challenges the notion of professionalism in hospitality. Sense of privacy and safety is valued by the guests The analysis suggests a positivity bias in Airbnb users’ comments while negative sentiments are mostly caused by ‘noise’. This research offers an alternative approach and more coherent understanding of the Airbnb experience. Methodologically, it contributes by illustrating how big data can be used and visually interpreted in tourism and hospitality studies.