An Experienced Business Developer vs A Lazy Traveler on AirBnb Seattle and Boston Data

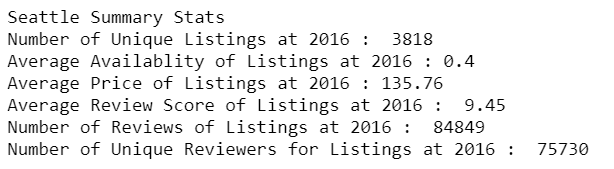
*A self-learner newbie’s dilemma on first DS blog post/homework project.*

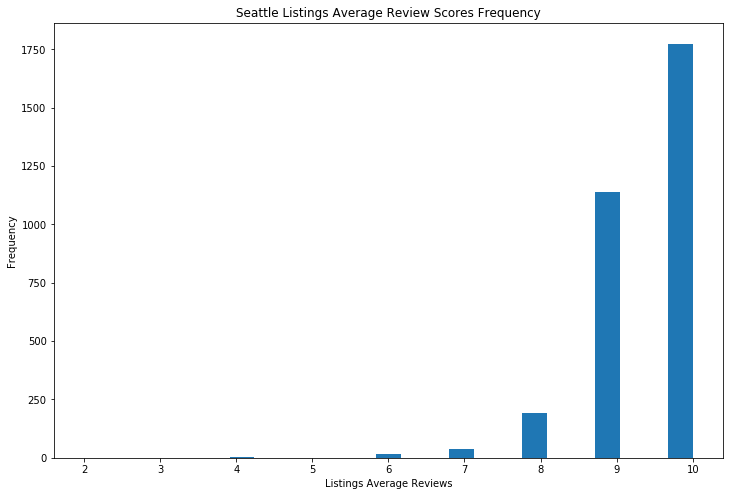
Taking Udacity’s Data Science Nanodegree program is a big challenge for me since I’ve believe in self-learning and data. As a business development professional nearly for a decade I’ve create and present countless reports, presentations and one-pagers (bosses of mine loves that ☺). On the other hand writing a data science blog post project homework had become one of the most difficult ones since making conclusions on a topic I’ve just started to learn is very terrifying.

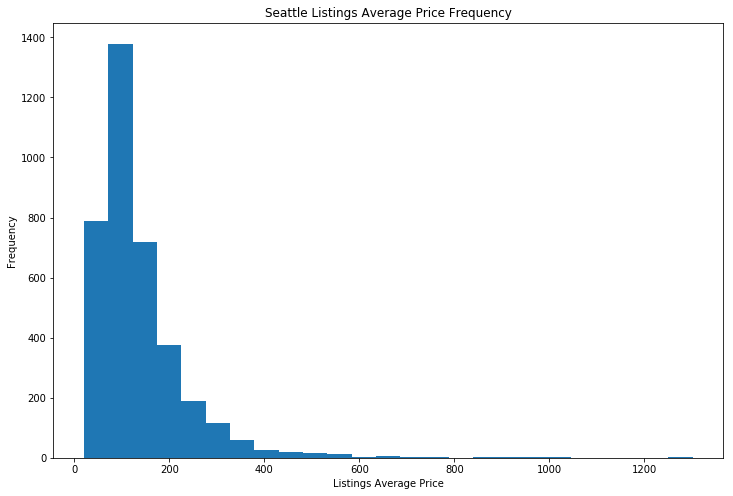
After ran some errands (cleaned the desk, updated windows os, watched a couple of episodes on Netflix) and heard ‘do your homework!’ yelling from my wife like my mom used to do 2 decades ago I’ve stated to analyze the given data sets for the project: Seattle AirBnb Data for 2016 and Boston AirBnb Data for 2017.

Each dataset contains 3 files as calendar.csv, listings.csv and reviews.csv. Listing contains every listed places on AirBnb with many features, calendar contains availability data of listings for 2016-2017 and reviews as the name suggest reviews written for the listings.

Below are some summary statistics for the Seattle Data which I’ve chosen to analyze first.



It seems there are enough data points namely listings on data set to make meaningful analysis. It seems 3818 listings have average review score of 9.45 out of 10 and average price for a night stay in Seattle is 135.76 US Dollars. Also there are more than enough reviewers for listings over 75K. A closer look on average reviews and prices below here:



Average prices histogram is consistent with our summary statistics. But more interestingly it seems most of the reviewers opted to give high scores for their stay at listings. This makes me wonder what might actually reason and what to look first. As mentioned earlier I’m a business development professional and one of the many things I’ve learned is that for any business/project is perception and one of the best way to improve a business’s perception is good story telling.

**Maybe listings on AirBnb intentionally or unintentionally drives reviewers to give higher scores with more positive and better descriptions of their listings?**

To analyze this part my coding approach was very simple. I’ve joined all descriptive columns on listings.csv and merged every reviews from reviews.csv, tokenize them (removed punctuation, stopwords and lemmatized) according to listings ids. Then, I joined to every description and reviews to create below word clouds:

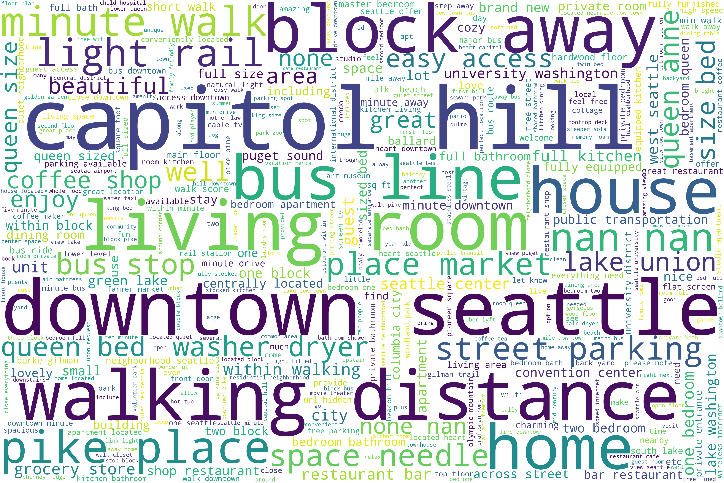
First one is word cloud of every listings description combined. It seems most used words in descriptive columns related with location and transportation. Also there are some word for bed sizes or nearby markets or coffee shops.

Figure: Listings Description Word Cloud

Also when it comes to reviews word cloud most of the highly reoccurring words related with recommendations. But also location related phrases are amongst highly frequent ones.

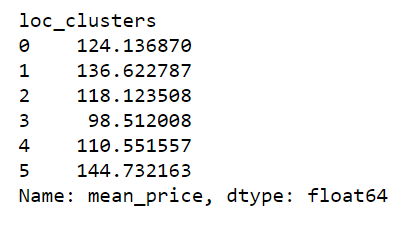
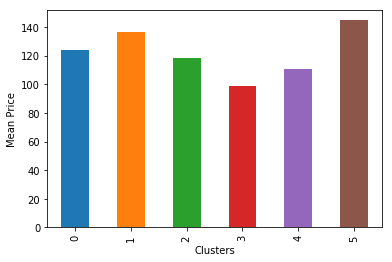
Figure: Listings Reviews Word Cloud

It seems my first question on is description wording is effecting reviews concluded as negative despite my high hopes. On the other hand failure of business developer has awoken lazy traveler inside me: ‘*See!! you always look location vs price when travelling and it seems you are not the only one.*’ This debate inside my head created the second point of my investigation:

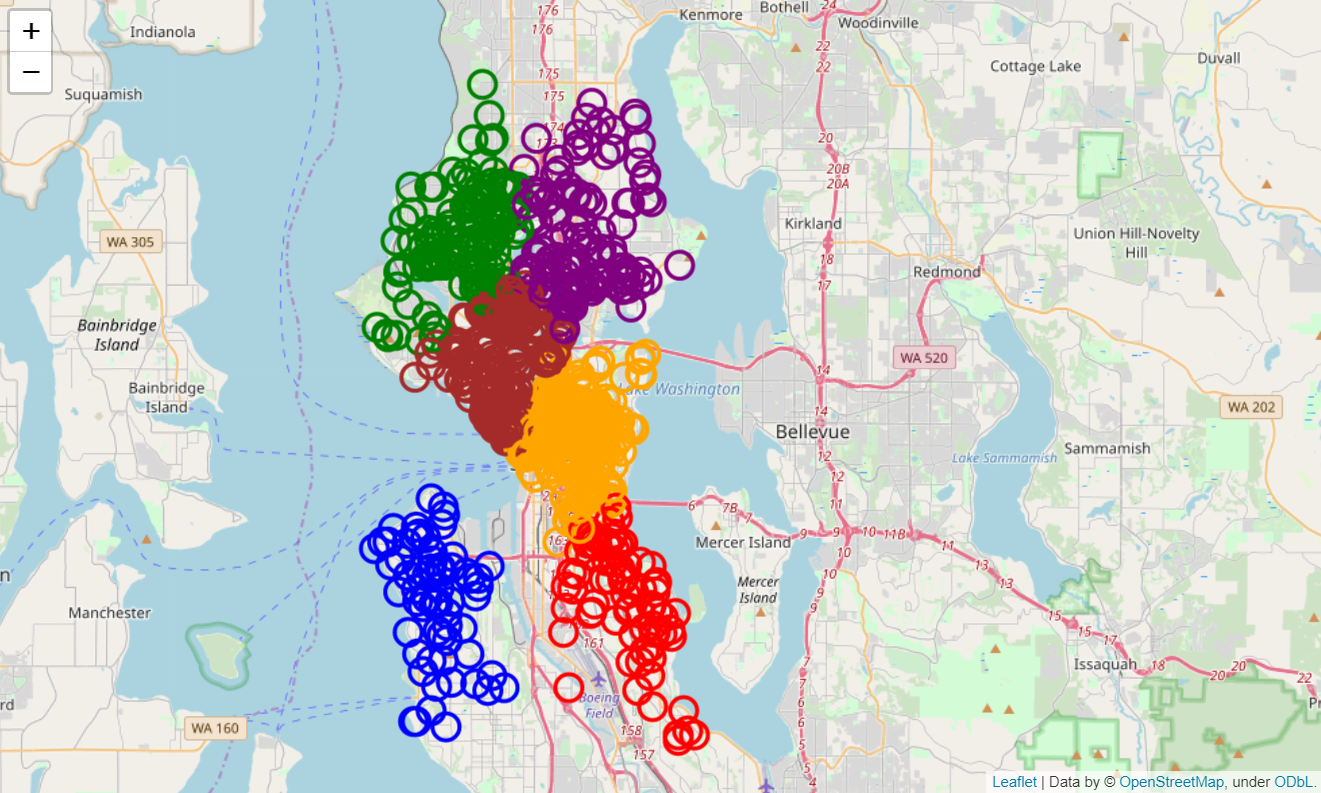
**Is location still has the biggest effect on any host businesses and their prices?**

To test this question again thinking and coding approach is simple. From calendar.csv I calculated average price per night for every listing ids. To remove outliers take listings between 2.5% and 97.5% relative to average prices confidence interval. Then from listings csv., selected ‘is\_location\_exact’ column True and take id, latitude and longitude columns joined average prices and dropped null values. Finally I had a data frame consisting id, latitude, longitude and average yearly price of the listing.

On this data frame I’ve applied to KMeans clustering only location data (average prices not included to KMeans) with cluster number equals 6. Below summary stats of predicted clusters according to average prices of every cluster:



It seems with some small and significant differences on average prices we had 6 different clusters. But the real magic happened when I’ve placed a thousand listing random sample and color map their clusters on the Seattle map with folium python library.



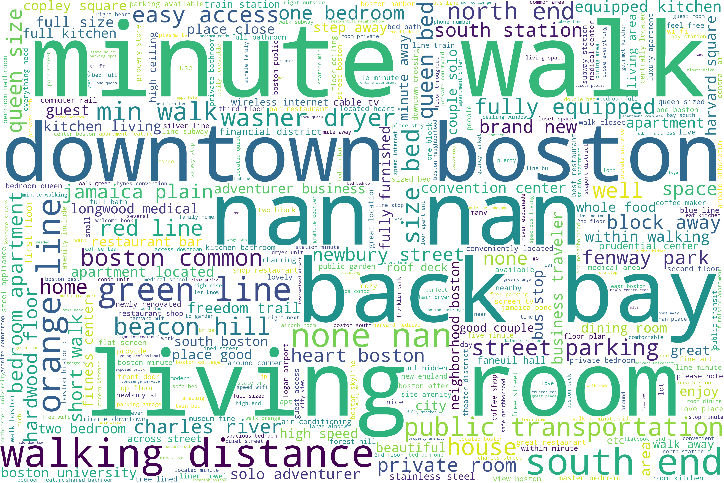
I wasn’t expecting that results that clearly. It seems our clusters according to coordinates of listings and their average prices per night matches with the idea of prices decreases from city centers to outside. Cluster numbers 5 and 1 which have highest average prices are located on downtown Seattle. 4 other clusters are located north and south of the downtown.

By the help of the map and summary stats on clustering we have made on listings on Seattle we can conclude that host businesses on city centers have higher average per night prices therefore high revenues relative to the outer ones. My second investigative question safely replied as yes good location equals higher revenues.

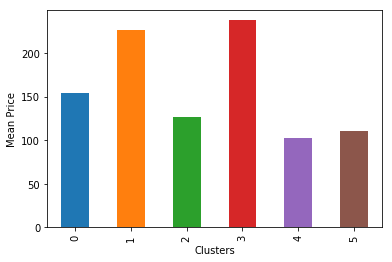
**Last and final point to discuss on this data set is my above findings are coincidental or can I prove these points on another city’s data?**

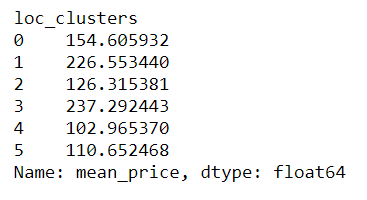
To test this point I’ve applied above steps on AirBnb Boston data which have same information points like Seattle data with only one difference the year. Boston data covers listings in 2017 rest of the data gad same structure and it was easy to implement above steps on the data.

First word clouds of Boston data are presented below:

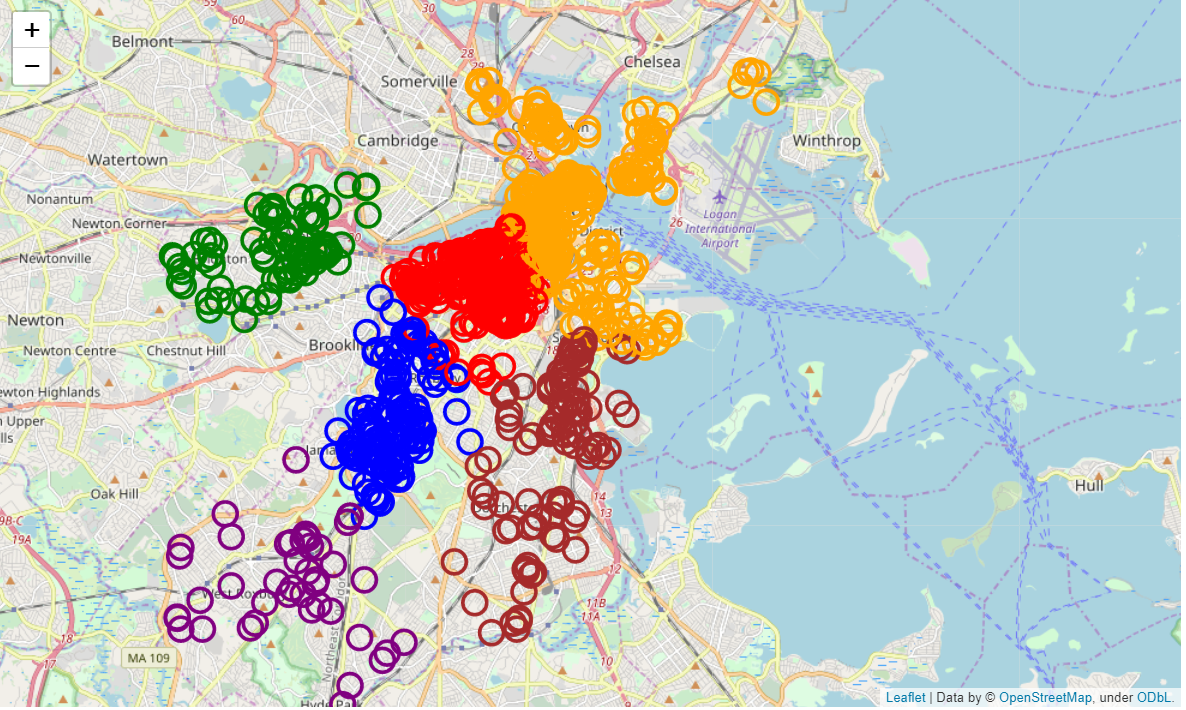
Figure: Boston Data Descriptive Columns Word Cloud Figure: Boston Reviews Word Cloud.

Again it seems that my first question on positive description effects reviews failed since most used words don’t match. On the other hand it is still encouraging to test good location results better revenues theory.

Applying same steps for data below summary statistic of our KMeans algorithm predicted clusters:



There are significant differences on average prices on our clusters based on coordinates. Below is the 1000 listing sample located on Boston map colored according to their clusters:



Magic happens again. It is clear that host business on central location have higher average prices and on the long run will have higher revenues.

To sum up, although my initial question on use of language effects reviews has failed according to most frequently used words on these to data sets, good location of any host businesses, city centers on our case, results higher revenues.

Finally these results are findings of a newbie on DS and created for a homework project. One can spend much more time on the data and can test words more in-depth for example effects of positive adjective or predict clusters not just coordinates but also actual distance to some city landmarks, transportation hub. As it comes for me my main take away from this project is lazy travelers knows best.

Happy travels.