DeepDive into Air BnB listing Dataset for Udacity Nanodegree program

This is my first project as a part of Udacity Nanodegree Data Science course. For this project I have selected Seattle Air BnB listing dataset. You can find the data in Kaggel <https://www.kaggle.com/airbnb/seattle/data>.



***Picture credit Wikipedia***

Before I start this blog, I would like to thank Udacity for their wonderful course material and prompt guidance by the mentors.

This blog aims to walk you through the project motivation and the steps I followed to derive the result.

As suggested by Udacity, I followed the CRISP-DM process(Cross-Industry Standard Process of Data Mining), which has 6 major phased, 1) Business Understanding, 2)Data Understanding 3) Data Preparation 4)Modeling the Data 5) Evaluating the results 6)Deployment.

In this project, I tried leveraging this opportunity to clean the dataset by following the process suggested by Udacity. I started with dataset having 3818 rows and 92 columns, after cleaning we have 3818 rows and 65 columns.

I didn’t want to go through the drop column route much as that is not a best practice, however I deleted some which I found absolutely unnecessary for this analysis or for any of my future analysis

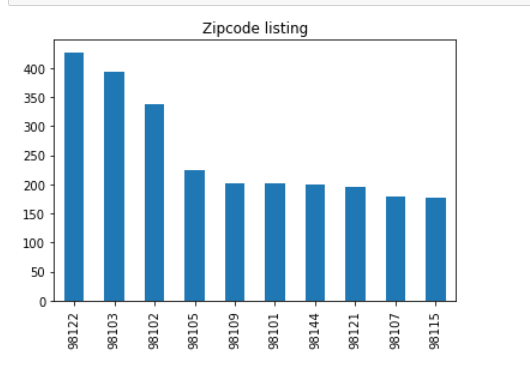
You can find code and detail analysis in **Github Repository**

**Project Motivation**

The motivation behind this project is to find answers for the following questions:

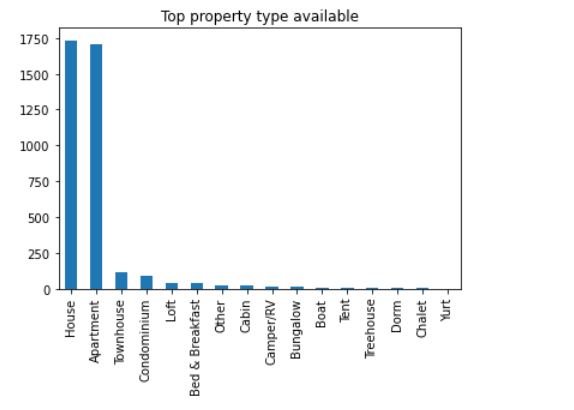
1. Top Zipcodes with most number of listing?
2. What kind of property is mostly available?
3. What are 10 most popular amenities?
4. Which review score is most important that impacts rating?
5. **Top Zipcodes with most number of listing.**

To find the Zipcode that has most listing I extracted number of unique Zipcode from the dataset. From our analysis we can say that Zipcode 98122 is the most frequently occurring Zipcode. Using this information AirBnB can plan promotion activity in the area where number of listing is low.

****

1. **Most Available Property Types**

To get the property type that is most available, I extracted number of unique property and found that most available properties are House and Apartments**.**

****

From the visual itself it looks like AirBnB Seattle has more than 90% listed properties are either House or Apartment.

1. **Top 10 most popular amenities**

To find the 10 most popular amenities that could be found in any AirBnB listing is as shown in the graph below. To derive this I created a unique list of all amenities using excel text to column feature and the removing the duplicates. This way we got unique list of amenities. Then I used a function(provided by udacity) which looks into each row and finds the number of times a particular amenity appeared in the dataset .

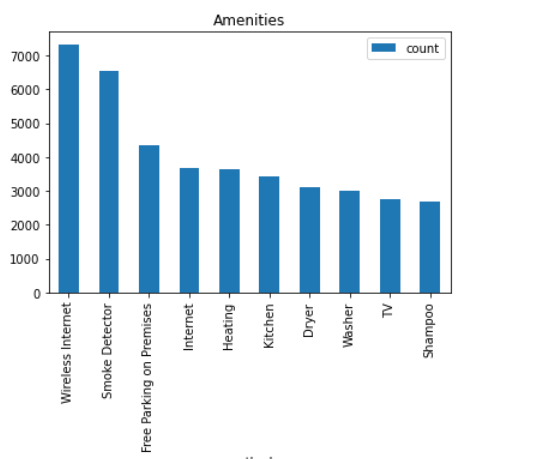
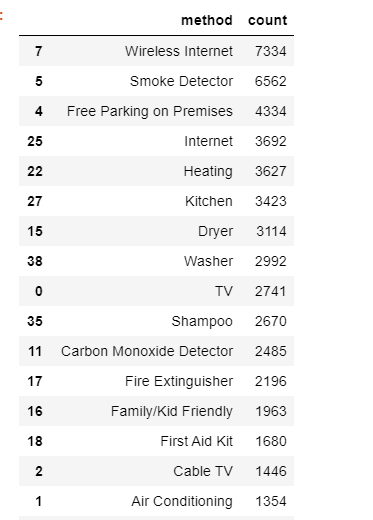
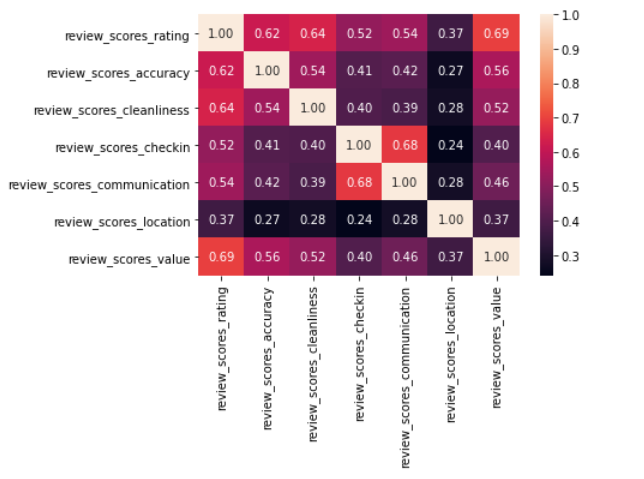
****

Figure List of Amenites in decreasing order



1. **Review Score that impacts rating**

To get the correlation between different review scores, I created the below heatmap . From the heatmap we can say that review Score value has most impact on rating.



At the end of the project I found the accuracy of the model. The model that I developed can explain variability upto 2%. You can find this is --------

**Conclusion**

In this project I have mostly done exploratory data analysis. Using this cleaned data we can identify many more patterns and insight in the data. I will continue to do my analysis and update this space gradually.

In the next project, I will try to do deeper analysis of the data and come up with more meaningful insights.

Thanks once again to Udacity and my mentor for their guidance in up skilling me.