Data Application Development Project

(Airbnb Listings in London and New York)

Jack Gallaher  
*BSHC Computing (Software Development)*  
*National College of Ireland*x16347436@student.ncirl.ie

1. Executive Summary

For our Data Application Development module, we have been tasked with analyzing and interrogating 2 or more datasets. They must be related to each other in some way. They both must complement each other and must be reasonably sized. For my Data Project, I have chosen to talk about the topic of Airbnb Listings in two major cities for tourists. Airbnb in London and New York City. Within these 2 datasets they give descriptions of the all Airbnb Services based on both cities. For my Project idea I have decided to contrast and compare certain values based off both the datasets. I will be comparing values of the Airbnb services in both New York and London. I will be comparing values such as Price of the services. The overall goal is comparing the values of both cities and see what is the difference between both. With two different cities there is bound to be a difference. I will also be manipulating data from both sets.

1. Project Setting

For this Project Idea as mentioned before I have chosen 2 datasets. The Datasets are based on Airbnb Services based in 2 major cities London and New York. Within these datasets There are certain characteristics that define the Airbnb’s services. With these characteristics I aim to retrieve values from certain Airbnb’s listed. One of my goals in this project is to compare certain values and headers about the Airbnb’s services in both cities and see which is better. Within these datasets some of the characteristics mentioned is price, which is mentioned in both datasets. One objective I would like to achieve in this project is comparing the price ranges from some of listed Airbnb services in both cities. By comparing them I would to like to see which is more expensive. with the price ranges listed from the services in both cities it could be based off the location of the Airbnb. Not just the city but the location from within the city.

I would also like to investigate the popularity of the Airbnb services. This can also be based on price and other attributes as well such as location, number of nights customer have stayed and how many reviews the place has received. over its duration of opening, these are not the only factors I wish to investigate, I would like to look into other studies as well. I would also like Manipulate some data from within the datasets by removing values and sorting out the datasets. I can show off the comparison studies by representing the results off the data that I compared in the form of graphs. In the form of graphs, it will make the data look more appealing and show the meaning of the results in more understandable way.

1. Literature Review

New York Dataset

For the Datasets I have chosen I must find other case studies that my datasets have been used with before. One of my Datasets I have chosen was the New York set. From my research gathered based on this dataset I have found certain discussions that my dataset has been used with before. Another user did an abstract on the data exploration of NYC Airbnb which is very similar to my study. He has stated that by using the data in my dataset he was able to discover that since 2008 guests and hosts have used Airbnb to expand on travelling possibilities. Today Airbnb has become a one of kind service. The data in the list can be analyzed and used for security and understanding of customers. Meaning that we can get a full understanding of what the customers really want when they search for the accommodation that suits they’re needs. Another study found about the New York dataset was entitled “hospitality in era for Airbnb”. Based on his study he used the data set to see how many people use Airbnb. With the statistics used he then found out that the dataset proves that Airbnb is used so much in New York. Another user used the dataset to compare R programming vs python programming. He did this by performing certain programming techniques on the dataset. Another user performed a sample activity about the number of Airbnb’s Reviews in New York.

London Dataset

During my literature review case findings, I also found other case studies based off the London dataset. The studies I found were quite small though. Other people created their Excel sheets and pasted certain attributes from the dataset into their own sheet. The studies were covered around the same period in the November 2019. The dataset was used to look for finding such as Detailed listings for London. For that dataset created they listed out just the listings of all the Airbnb’s in London. Other datasets that were created involved “detailed calendar data for listings” in London. The other datasets that were created off the London Datasets were looking at specifically reviews of the all the Airbnb services. Another was based on the “Summary information and metrics listings”. Finally, a set was created based on the reviews and listing Ids of the Airbnb’s. This set was used to “facilitate time-based analytics and visualizations linked to the listings”. I also discovered that through the use of the London dataset there was a visual representation of the city of London created with green and red dotes to show the availability of the Airbnb’s services throughout the city

1. Data

For my Chose of Data usage in the data application project I have chosen 2 datasets about Airbnb in London and New York. My reason for choosing the topic of Airbnb is because as we all know holidays can be an expensive thing to plan out. When it comes to flights, they can be quite expensive as well as accommodation which is an important factor. Hotel prices are ridiculously high everywhere. With Airbnb services introduced customers have found a much cheaper way with accommodation which gives them their privacy instead of using a hostel. Travelling and accommodation finding is big thing internationally with holiday planning. Airbnb has become very popular everywhere. It is estimated that they have over 4 million listings in over 65000 cities in 191 countries. I also feel that with the 2 datasets I have chosen, since they are both based off on Airbnb listings and they have the same attributes, the comparison section should be made easier and it means they can both complement it each. The datasets also complement each other because the data within both were recorded in the same year (2019). This is good as we can see from the same time period how each listing in both cities have done well.

When I was trying to come up with ideas for my project, I decided that in terms of trying to interest me I wanted to do something based off travelling since it really interests me so much. From places I have visited I really enjoyed London and New York. Since they are two very popular cities for visiting internationally, I figured that doing something on them 2 cities would be good.

|  |  |  |
| --- | --- | --- |
| File | Airbnb London Set (AB\_LON\_2019) | Airbnb New York Set (AB\_NYC\_2019) |
| Format | Excel Sheet | Excel sheet |
| Size | 11.172 Megabytes | 6.956 Megabyte’s |
| Number of Instances | 83851 Records | 48896 Records |
| Number of Attributes | 16 | 16 |

|  |  |
| --- | --- |
| Attribute | Type |
| id | Int |
| name | Factor |
| host\_id | Int |
| host\_name | Factor |
| neighbourhood\_group | Factor |
| neighbourhood | Factor |
| latitude | Numeric |
| longitude | Numeric |
| room\_type | Factor |
| price | Int |
| minimum\_nights | Int |
| number\_of\_reviews | Int |
| last\_review | Fator |
| reviews\_per\_month | Numeric |
| calculated\_host\_listings\_count, | Int |
| availability\_365 | Int |

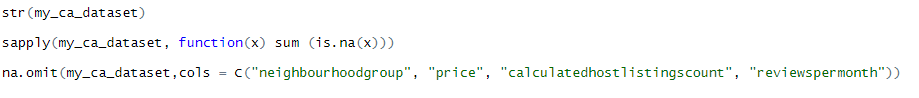
1. Methodology

KDD Methodology

For our project we had to use 1 of the following methodologies KDD or CRISP-DM. For this project I followed the steps of the KDD methodology. For this process I performed steps that involved the following. The first step in KDD is selection. For this step I had to retrieve the dataset. The first dataset I retrieved was the New York Dataset from Kaggle.com This Dataset was downloaded through the following website as an excel sheet. The 2nd dataset was the London dataset was retrieved from the Data.world.com. It was also downloaded as an excel sheet.

The next step is preprocessing (Target data) the data from the set. For this step I wish to choose certain attributes to work off from. I do this so I can perform some kind of analysis on the data for a small case study. I wanted to specifically to talk about prices of the Airbnb’s in both cities. I also wanted to talk about the popularity of room types that customers wanted to choose. I performed this task for both datasets as the datasets are both similar to each other.

Within these the Datasets I attempted to clean the data. To perform this, I used R (Programming Language) to clean the data. I first had to summarize the data by retrieving all the attributes. This would return the output by showing all the attributes from within the dataset showing what attributes had missing values. I would enter a piece of code that would contain the attributes that having missing values and should remove the columns from within the dataset that would have missing values.

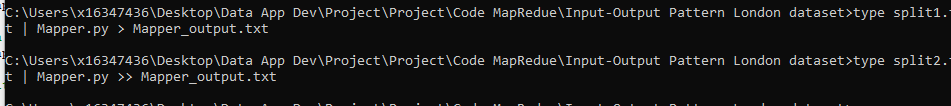


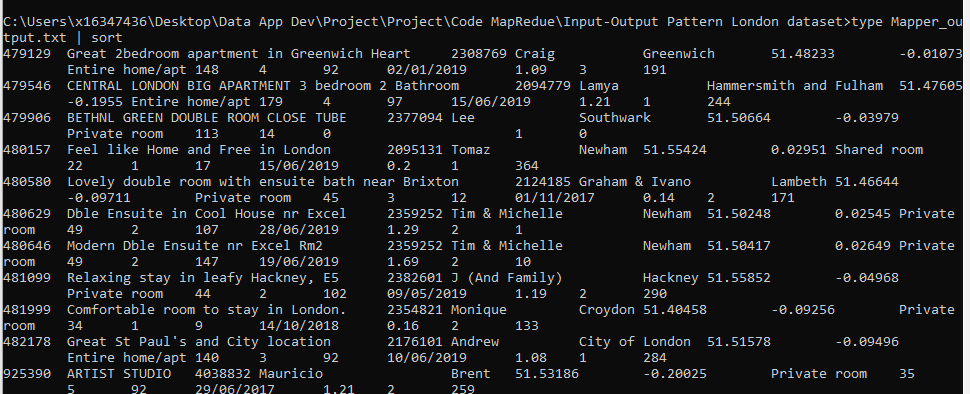
the following code was supposed to retrieve the dataset headers, followed by finding the missing values which was successful but the final line code which was supposed to remove the columns with N/A values did not work. It worked the same way for the both datasets as they both had the same attributes (headers). so, I performed the same code on the both datasets when I was trying to clean the data.

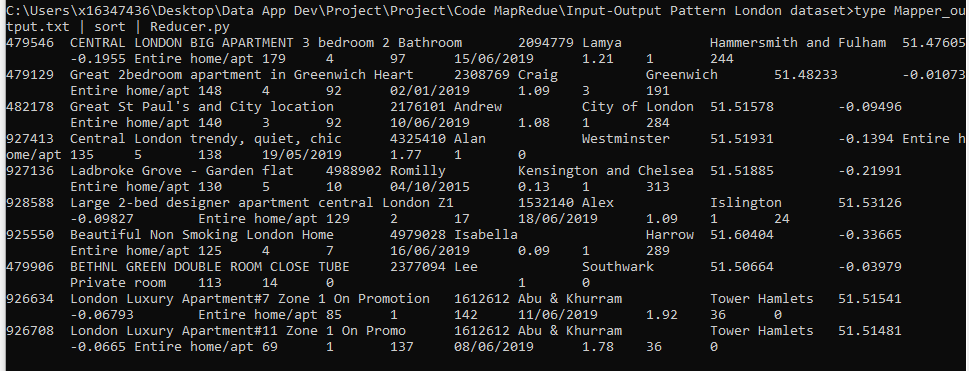
The next step involved in the KDD methodology is transformation. Within this step we have to transform the data. This step involves us manipulating the dataset and performing functions. This involves us performing functions by representing data in the form of graphs and following the MapReduce procedure. This is how I show off my analysis.

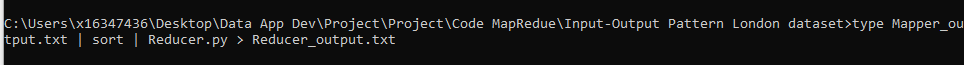
The next process where I used KDD Methodology was when I performed the MapReduce Patterns on the datasets. The first MapReduce pattern I did was Input/output. I followed the same steps as before mentioned. The selection step was already complete as I had downloaded the datasets already. For the preprocessing step (target data) I selected a numerous amount of records from both datasets (2000 from each dataset). For each set I copied the 1000 out of the 2000 and pasted 1000 into a text file. Both named Split1.txt and Split2.txt for both datasets.

The next step was transformation which involved performing so kind of manipulation with the data. For this I ran the mapper file for both Split1 and Split2 text files on both datasets. This then mapped out certain records that were considered important. They were mapped into file called Mapper\_output.txt. I then performed the sort function which will sort out the records within the files. I then had to perform the reducer method on the mapped data, followed by performing the sort function again, which will sort out the records within the files. This then broke down the mapped records in tuples. The remaining records for both datasets were moved into a file called Reducer\_output.txt. the final step involved me moving the reducer file outputs from both sets into a file called Combinedsorted.txt where both London and New York mapped values were joined.







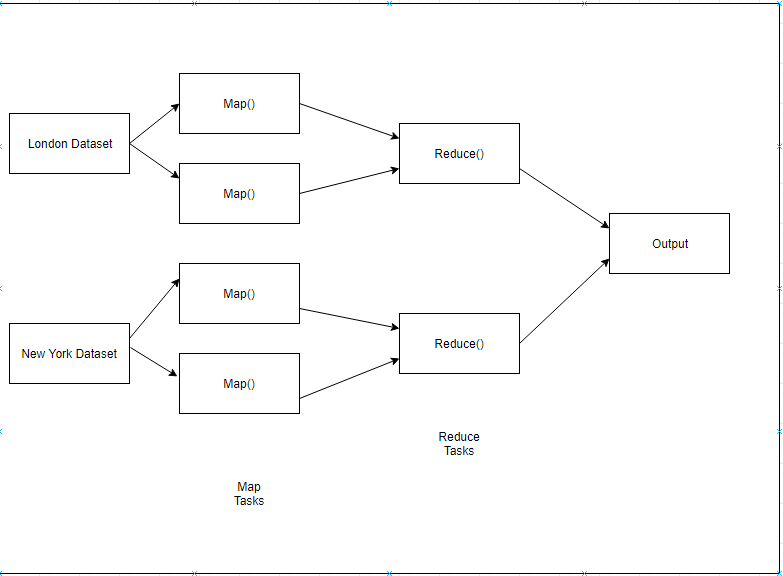


1. Implementation and architecture

For the workflow of how my data application project works I performed my data analysis study using the Graphs as a representation of what I wanted to see from my results. I performed this using R in R studio. With this I ran my code which would then output the results of analysis as a visual representation. I checked for certain variables such as price. The comparison of these values would be outputted as graphs. For my MapReduce patterns I used Python to create the mapper and reducer files to perform those functions such as input/output and summarization. I performed the execution of MapReduce using the command prompt. The code for python was stored in NotePadd++. In terms of automation, I would say that the project is not entirely automated. With the R studio Functions to retrieve the graphs you must run each analysis individually to retrieve an output. For example, to get a graph for price vs minimum nights you would run that first. Then run the next line of code to get another result.

To run a python function for MapReduce you must run each command individually to get an output. For example, you must the mapper first. Then run the reducer after.

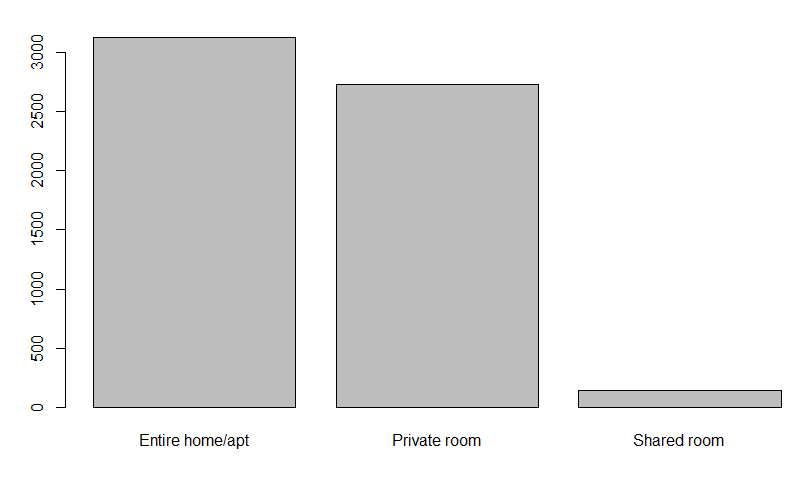
MapReduce Diagram:

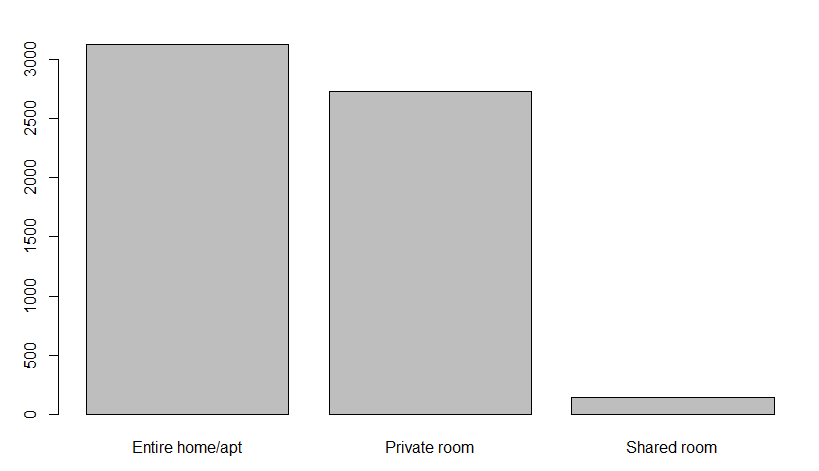


1. Results

For this section I will discussing my analysis case study that I wanted to perform for the project. As mentioned in the introduction I wanted to perform certain studies based off both datasets. My overall goal for this study was look at certain attribute values from both cities and compare them. Within this section I will show my analysis results by which I gathered using graphs. The graphs were created using programming language R

The first study I looked at was the room type comparisons in both London and New York. This section discusses which type of room that all customers in both cities preferred

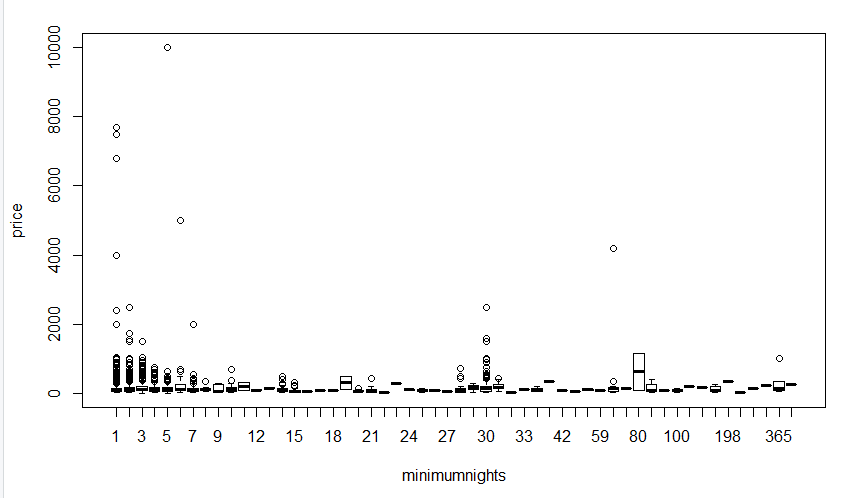
New York/ Room Types



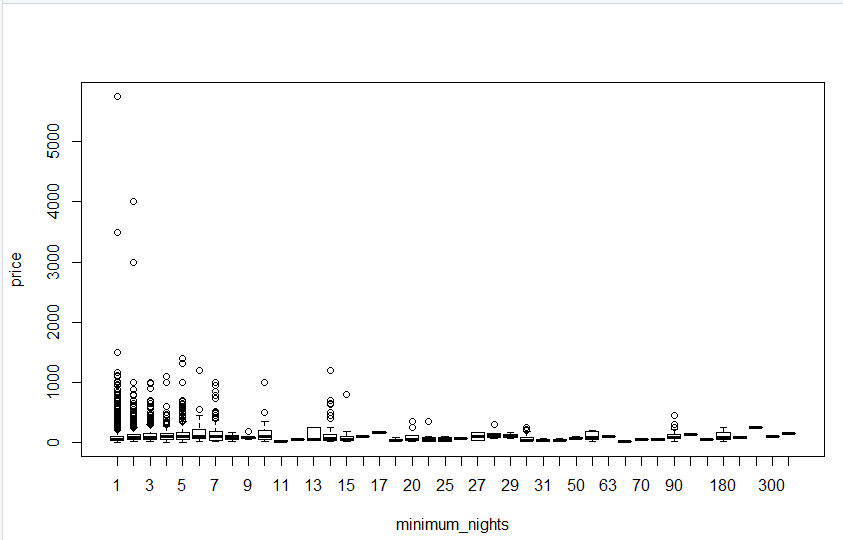
London/ room Type

As you can see from both graphs, they are both very similar. This insight discovered tells that most of the customers who use Airbnb in New York and London prefer to rent out a listing that is an entire home/ apartment. This can be related to the price of the listings and also location. Another reason would be because of the prices of private rooms and also having to share rooms with other people.

Another insight that I found during my analysis was the price range versus amount of nights customer stay in the Airbnb.



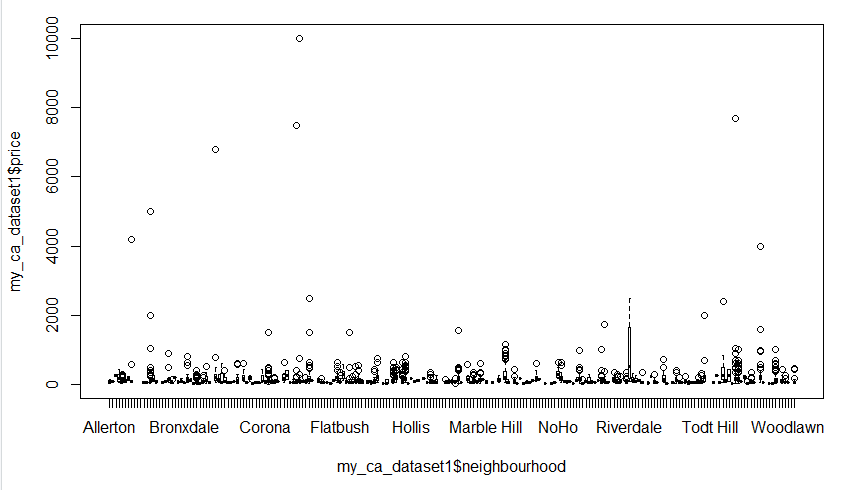
New York/ Price vs Minimum nights



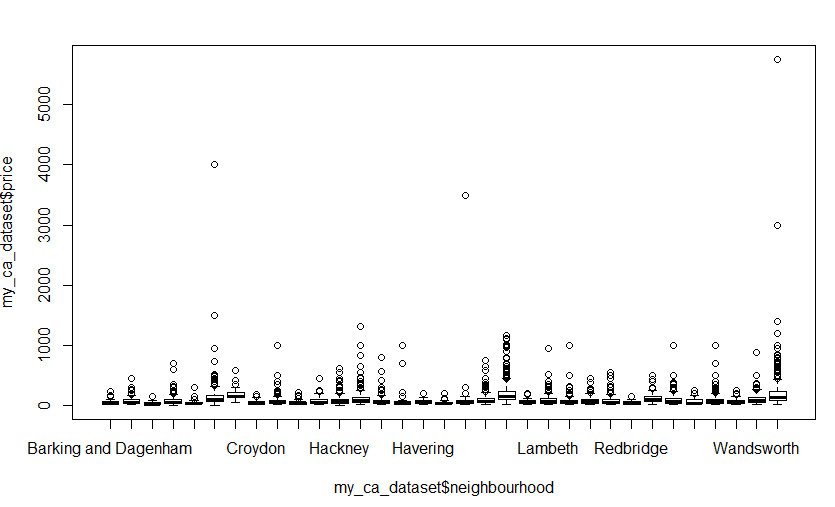
London/ Price vs Minimum nights

As we can see from this representation in the form of graphs. With both cities we can see that majority of customers stay from 1 to 7 nights minimum in both cities. We can also see from the price range in both cities that all customers will go for particularly low prices (in the hundreds or lower). It’s very rare when you see someone renting at a higher price. From the dataset most prices range from low to hundreds and this shows us that most people stay a few nights at a relatively low but reasonable price.

The next insight I saw was the neighborhood versus the price.



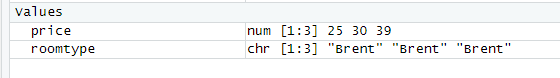
New York/ price vs neighborhood



London/ price vs neighborhood

From the dataset I also retrieved graphs based on the price vs the neighborhood. These are not all the neighborhoods in both cities but some of the major areas. As we can see the prices are relatively similar in some areas within both cites. When it comes to price it really depends on the part of the city you stay in. Inner city can have high prices and outer parts can mostly be low because of common knowledge we know that the inner city is a more tourist area with more attractions to do so therefore it becomes a tourist trap for visitors.

From both datasets I gathered some data for comparison of prices to easily explain my last point on prices based off location.



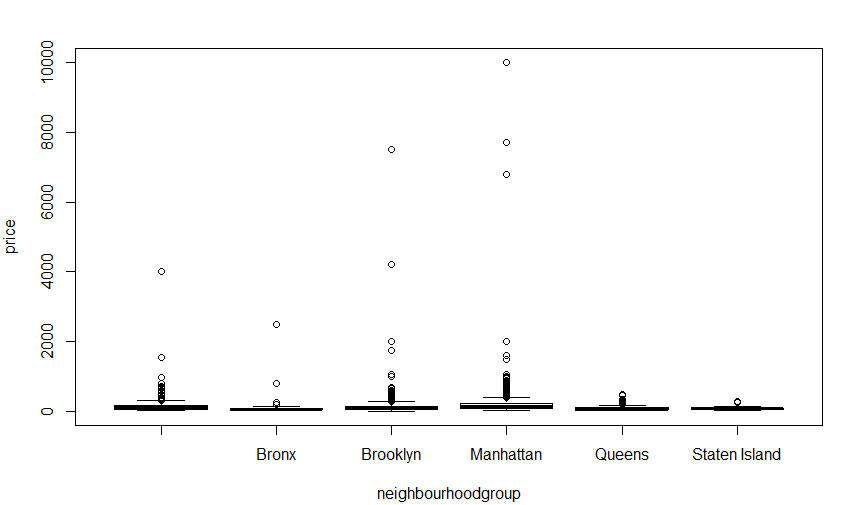
As you can see from these 2 containers above, I created based off London, we can see Westminster prices are high as its based in the center of the city. Brent, which is outer London based, has lower prices. As we know Westminster other then it being city center based it’s a very popular tourist area.

As we know cities are broken down into districts. With London there are 33 districts. Within the dataset for London we see that the attribute neighborhood group is empty as it differs from New York.



In the New York Dataset neighborhood group, we can see that the 5 main districts of New York are listed, Manhattan, Brooklyn, Queens, the Bronx and Staton Island. This is because neighborhood group differs in both cities.

One Last insight I found with my data comparison was in the New York Dataset I noticed that prices are mostly evened out in the 5 districts of New York. But as we can see Manhattan has slightly Higher prices compared to the other 5. This would be a result of Manhattan being a very popular district for visitors.



New York/ Price vs Neighborhood group

1. Conclusions and future work

From my results gathers from this analysis I conducted it has shown me the differences in both cities with prices of the Airbnb listings. I have seen how in both London and New York the prices can be high, depending on the location of the listing. I have seen how the majority of customers prefer to have an entire apartment/house to themselves rather than the other options listed. I have also discovered that in terms of which city is better. There is not much of a difference.

As a result of this Data set project, it has taught more about the use of R programming. It has taught me a lot about the how to extract data from data sets and ways of how to perform data analysis of particular values based within the dataset. It has also taught me how to view visual representations in the forms of graphs on particular aspects of datasets. On the topic of my 2 datasets it has shown me the differences of accommodation usage in both cities. I have also learned about the popularity of Airbnb. In the case MapReduce, I have to learned to map data (taking sets of data and converting it into another set of data). I have learned the aspect of reducing the data (by breaking it down into a smaller dataset with just the important elements and getting rid of the rest).

If I was to go back and do this project again. I would perform different MapReduce patterns with the dataset, instead of performing the Character counting and input/output techniques I would try other functions with the dataset such as retrieving the max/minimum values of the dataset or try to find the occurrence of the amount words from within the dataset. I would also when using the Programming language R would to try compare other values from within the dataset and perform a different analysis.

1. References

Datasets:

1. London Airbnb Dataset:

Data.world. (2019). *Inside Airbnb London - dataset by kgarrett*.

[online]Available at: <https://data.world/kgarrett/inside-airbnb-london>

[Accessed 20 Dec. 2019].

1. New York Airbnb Dataset:

Kaggle.com. (2019). *New York City Airbnb Open Data*.

[online]Available at: https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data

[Accessed 20 Dec. 2019].

Literature Review:

New York Dataset

Kaggle.com. (2019). *Data Exploration on NYC Airbnb | Kaggle*.

[online] Available at: <https://www.kaggle.com/dgomonov/data-exploration-on-nyc-airbnb>

[Accessed 21 Dec. 2019].

Kaggle.com. (2019). *Hospitality in Era of Airbnb | Kaggle*.

[online] Available at: <https://www.kaggle.com/biphili/hospitality-in-era-of-airbnb>

[Accessed 21 Dec. 2019].

Kaggle.com. (2019). *Python Programming versus R Programming | Kaggle*.

[online] Available at: <https://www.kaggle.com/ekrembayar/python-programming-versus-r-programming>

[Accessed 21 Dec. 2019].

London Dataset

Inside Airbnb. (2019). *Inside Airbnb: London. Adding data to the debate.*.

[online] Available at: <http://insideairbnb.com/london/>

[Accessed 21 Dec. 2019].

Inside Airbnb. (2019). *Inside Airbnb. Adding data to the debate.*.

[online] Available at: <http://insideairbnb.com/get-the-data.html>

[Accessed 21 Dec. 2019].