CIS5200 Term Project Tutorial

Authors: Alekhya Raidu Bojja Venkata, Ashima Ashima, Upma Kumar

Instructor: Jongwook Woo

Date: 12/09/2017

Lab Tutorial

Alekhya Raidu (abojjav@calstatela.edu)

Upma Kumar (ukumar@calstatela.edu)

Ashima Ashima (aashima@calstatela.edu)

12/09/2017

Airbnb Data Analysis using Hive

(Analysis on Airbnb Data)

Objectives

Following are the tasks that we are going to perform in this lab tutorial

- Download dataset and create Hadoop cluster
- Upload files in Ambari
- Hive commands to perform the data analysis.
- Visualization

Platform Spec

- Cluster Type: Hadoop IBM Bluemix BigInsights
- CPU Speed: 2.25 GHz
- # of CPU cores: 2
- # of nodes: 2
- Total Memory Size: 1TB SATA
- Data Storage: 24.4 GB

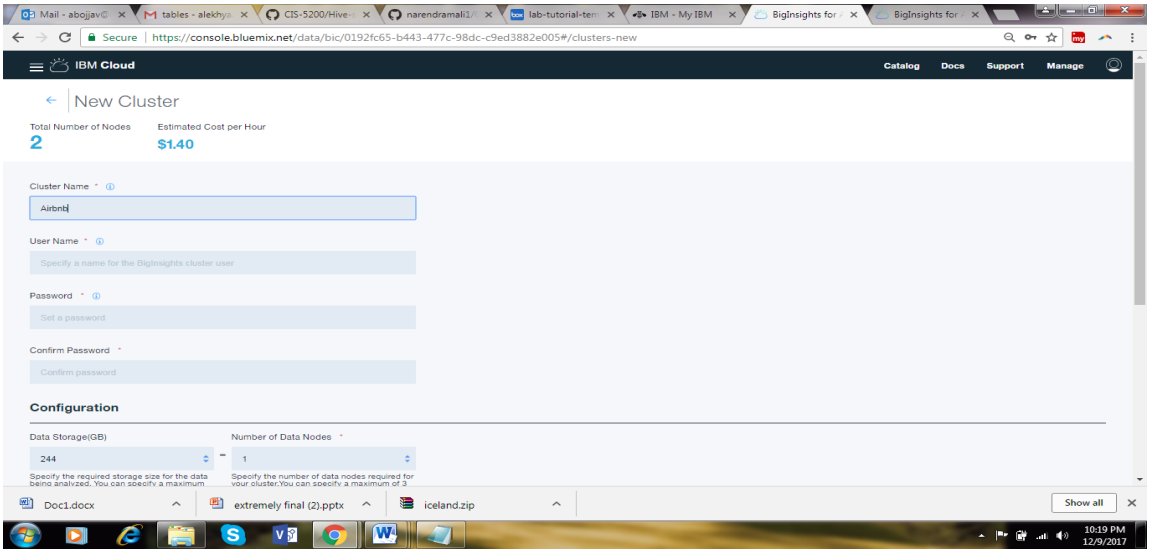
Step 1: Download Dataset and create Hadoop cluster

In this step, we are going to download dataset and set up a hadoop cluster.

1. Download dataset using the URL: http://tomslee.net/airbnb-data-collection-get-the-data

2. Sign into your ibm bluemix account

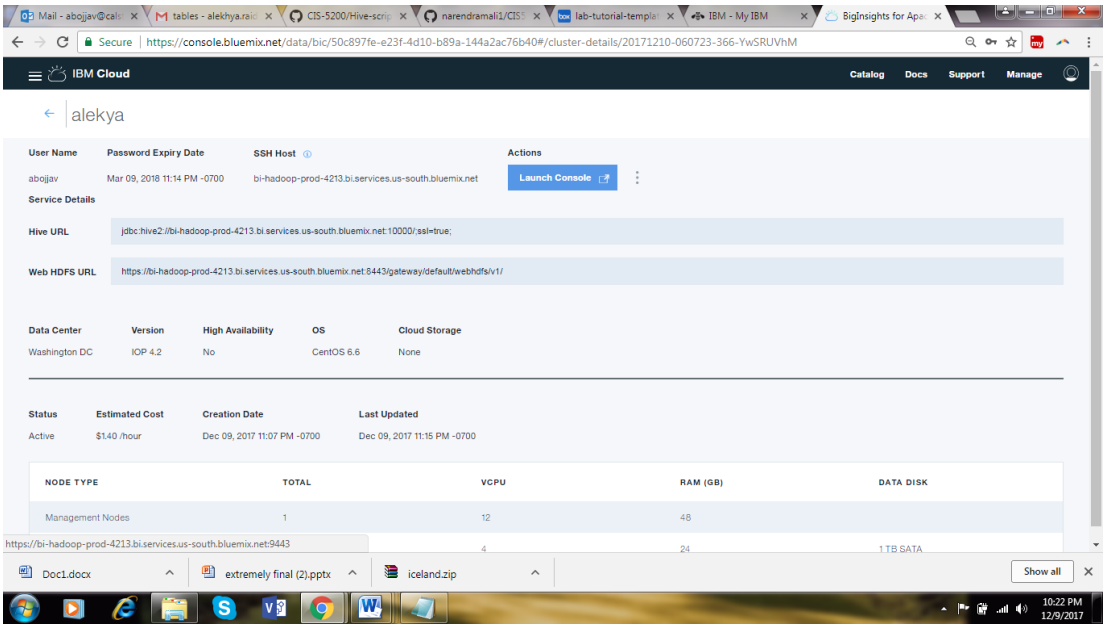
3. Create Hadoop cluster



Step 2: Upload files in Ambari

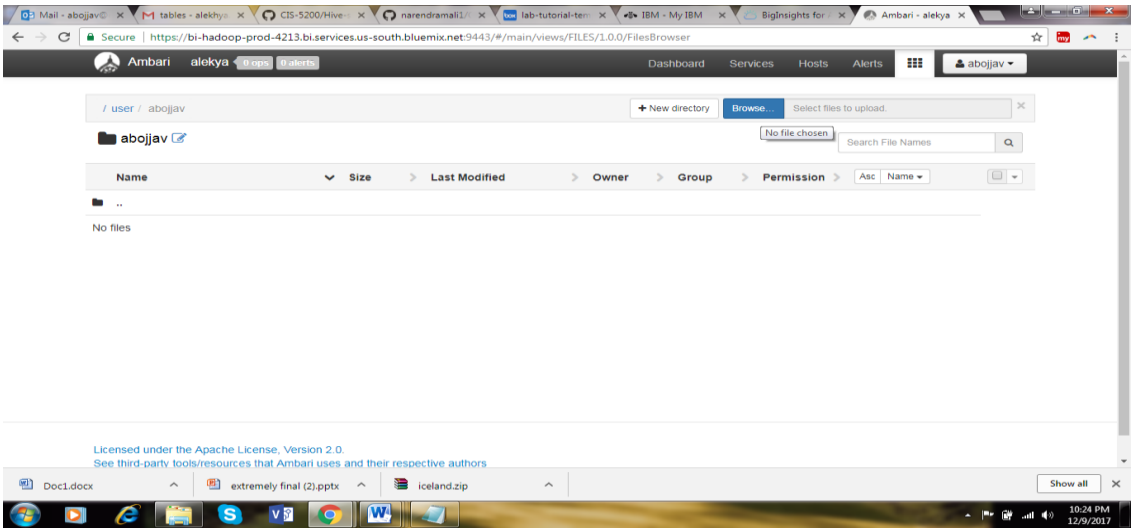
In this step, we have to upload files manually to Ambari

1. Open Ambari by clicking the Launch console in IBM blue mix account.



2. Sign in to Ambari by giving the same username and password at the time we have created the cluster

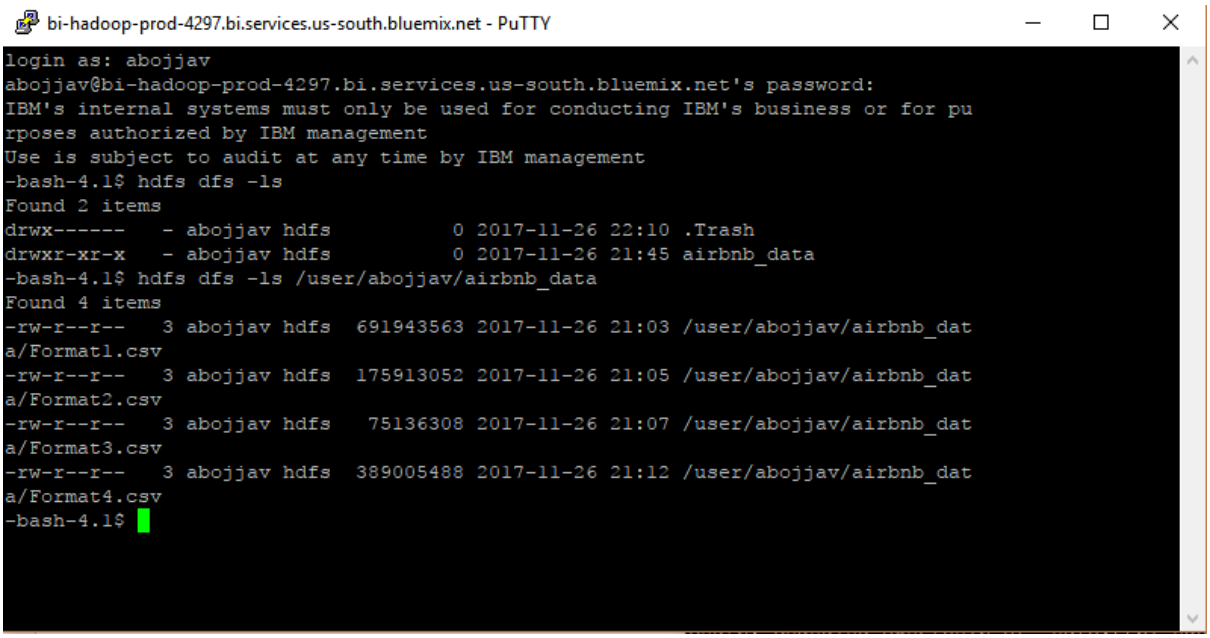
3. Go to file browser/user/abojjav/airbnb_data and click on upload as shown in the figure



4. Open Putty and giving SSH of Hadoop cluster and given Ambari's user id and password

5. Finally check the whether the uploaded files are there in folder or not using the following command.

```
-bash$: hdfs dfs -ls
```

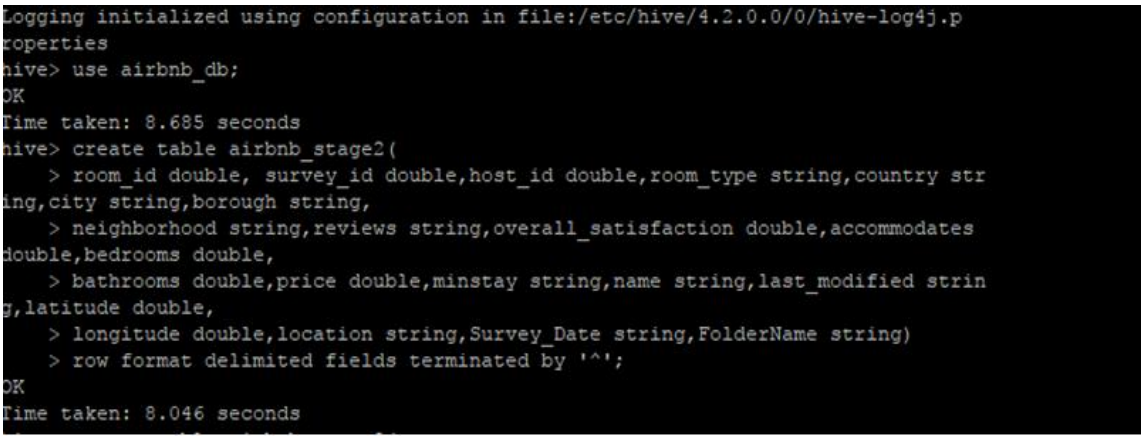


Step 3: Hive commands to perform data analysis

Here we are going to create tables for querying

1. Open Hive in command prompt of Putty
2. Use airbnb_db for querying data

```
Hive> use airbnb_db;
```



3. Create table for the first csv file which includes data of 10 cities

```
create table airbnb_stage1(room_id double,host_id
double,room_type string,borough string,neighborhood
string,reviews string,overall_satisfaction string,accommodates
string,bedrooms double,price double,minstay string,latitude
```

```
double,longitude double,last_modified string,Survey_Date
string,FolderName string)row format delimited fields terminated
by '^' tblproperties("skip.header.line.count"="1");
```

4. Create table for the second csv file which includes data of 10 cities

```
create table airbnb_stage2(room_id double, survey_id
double,host_id double,room_type string,country string,city
string,borough string,neighborhood string,reviews
string,overall_satisfaction double,accommodates double,bedrooms
double,bathrooms double,price double,minstay string,name
string,last_modified string,latitude double,longitude
double,location string,Survey_Date string,FolderName string)row
format delimited fields terminated by
'^'tblproperties("skip.header.line.count"="1");
```

5. Create table for the third csv file which includes data of 10 cities

```
create table airbnb_stage3(room_id double, survey_id
double,host_id double,room_type string,country string,city
string,borough string, neighborhood string,reviews
string,overall_satisfaction double,accommodates double,bedrooms
double, bathrooms double,price double,minstay string,name
string,property_type string,last_modified string,latitude double,
longitude double,location string,Survey_Date string,FolderName
string) row format delimited fields terminated by '^'
tblproperties("skip.header.line.count"="1");
```

6. Create table for the fourth csv file which includes data of 10 cities

```
create table airbnb_stage4(room_id double,survey_id
double,host_id double,room_type string,country string,city
string,borough string,neighborhood string,reviews
string,overall_satisfaction double,accommodates double,bedrooms
double,bathrooms double,price double,minstay string,last_modified
string,latitude string,longitude double,location
double,Survey_Date string,FolderName string)row format delimited
fields terminated by
'^'tblproperties("skip.header.line.count"="1");
```

7. Loading data into tables

```
load data inpath 'hdfs:/user/abojjav/airbnb_data/Format1.csv into
table airbnb_stage1;
load data inpath 'hdfs:/user/abojjav/airbnb_data/Format2.csv into
table airbnb_stage2;
load data inpath'hdfs:/user/abojjav/airbnb_data/Format3.csv into
table airbnb_stage3;
load data inpath 'hdfs:/user/abojjav/airbnb_data/Format4.csv into
table airbnb_stage4;
```

```
hive> load data inpath 'hdfs:/user/abojjav/airbnb_data/Format1.csv' into table airb
Loading data to table airbnb_db.airbnb_stage1
chgrp: changing ownership of 'hdfs://bi-hadoop-prod-4297.bi.services.us-south.bluemix.net:8020/app
s/hive/warehouse/airbnb_db.db/airbnb_stage1/Format1.csv': User does not belong to hadoop
Table airbnb_db.airbnb_stage1 stats: [numFiles=1, totalSize=691943563]
OK
Time taken: 1.583 seconds
hive> load data inpath 'hdfs:/user/abojjav/airbnb_data/Format2.csv' into table airbnb_stage2;
Loading data to table airbnb_db.airbnb_stage2
chgrp: changing ownership of 'hdfs://bi-hadoop-prod-4297.bi.services.us-south.bluemix.net:8020/app
s/hive/warehouse/airbnb_db.db/airbnb_stage2/Format2.csv': User does not belong to hadoop
Table airbnb_db.airbnb_stage2 stats: [numFiles=1, totalSize=175913052]
OK
Time taken: 0.198 seconds
hive> load data inpath 'hdfs:/user/abojjav/airbnb_data/Format3.csv' into table airbnb_stage3;
Loading data to table airbnb_db.airbnb_stage3
chgrp: changing ownership of 'hdfs://bi-hadoop-prod-4297.bi.services.us-south.bluemix.net:8020/app
s/hive/warehouse/airbnb_db.db/airbnb_stage3/Format3.csv': User does not belong to hadoop
Table airbnb_db.airbnb_stage3 stats: [numFiles=1, totalSize=75136308]
OK
Time taken: 0.169 seconds
hive> load data inpath 'hdfs:/user/abojjav/airbnb_data/Format4.csv' into table airbnb_stage4;
Loading data to table airbnb_db.airbnb_stage4
chgrp: changing ownership of 'hdfs://bi-hadoop-prod-4297.bi.services.us-south.bluemix.net:8020/app
s/hive/warehouse/airbnb_db.db/airbnb_stage4/Format4.csv': User does not belong to hadoop
Table airbnb_db.airbnb_stage4 stats: [numFiles=1, totalSize=389005488]
OK
Time taken: 0.173 seconds
```

8. Creating external table

```
Create external table airbnb_master(room_id double,host_id
double,room_type string,borough string,neighborhood
string,reviews string,overall_satisfaction string,accommodates
string,bedrooms double,price double,minstay string,latitude
double,longitude double,last_modified string,Survey_Date
string,FolderName string)row format delimited fields terminated
by '\t'LOCATION '/user/abojjav/';
```

9. Creating master table which is union of all the above 4 tables.

```
select count (*) from (select
room_id,host_id,room_type,borough,neighborhood,reviews,overall_sa
tisfaction,accommodates,bedrooms,price,minstay,latitude,longitude
,last_modified,Survey_Date,FolderName from airbnb_stage1
UNION ALL
selectroom_id,host_id,room_type,borough,neighborhood,reviews,over
all_satisfaction,accommodates,bedrooms,price,minstay,latitude,lon
gitude,last_modified,Survey_Date,FolderName from airbnb_stage2
UNION ALL
selectroom_id,host_id,room_type,borough,neighborhood,reviews,over
all_satisfaction,accommodates,bedrooms,price,minstay,latitude,lon
gitude,last_modified,Survey_Date,FolderName from airbnb_stage
UNION ALL
selectroom_id,host_id,room_type,borough,neighborhood,reviews,over
all_satisfaction,accommodates,bedrooms,price,minstay,latitude,lon
gitude,last_modified,Survey_Date,FolderName from airbnb_stage4)
abc;
```

10. Finding highest number of accommodate for each root type

```
Drop table if exists highaccommodates;
create table highaccommodates as
select count(accommodates),room_type
from airbnb_stage2
group by room_type;
```

11. Finding average overall satisfaction of Airbnb service in all cities

```
Drop table if exists satisfaction;
create table satisfaction as
select city,avg(overall_satisfaction)
from airbnb_stage2
group by city;
```

12. Finding reviewd and survyed cities in differet countries

```
Drop table if exists review;
create table review as
select country,city,survey_id,reviews
from airbnb_stage2
group by country,city,survey_id,reviews;
```

13. finding number of rooms booked in each city

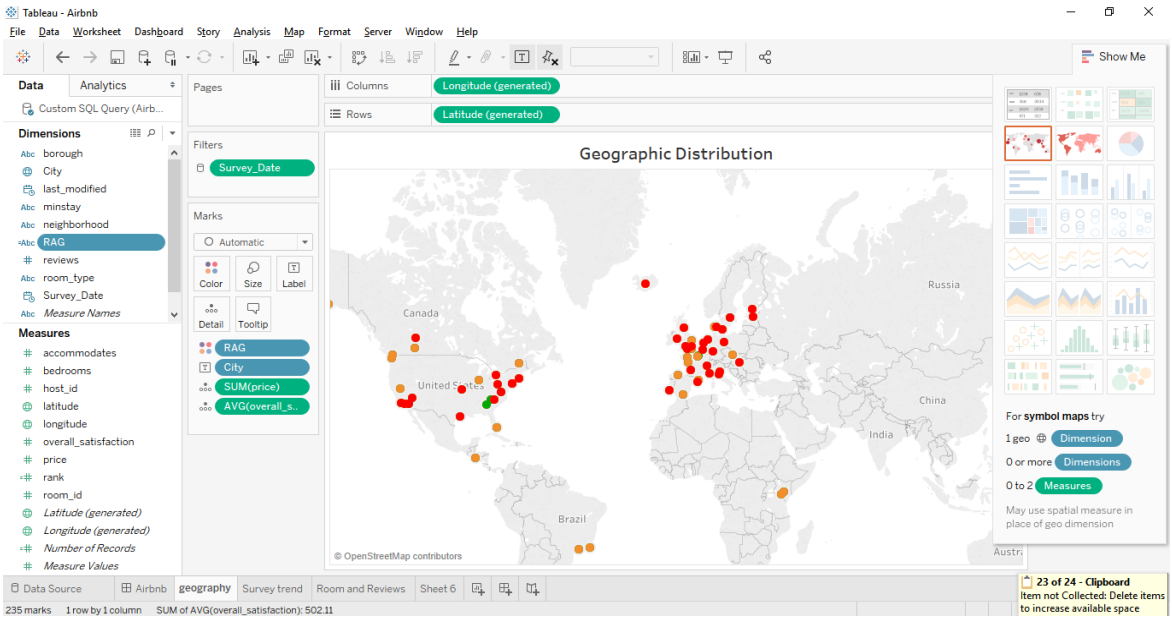
```
Drop table if exists rooms;
create table rooms as
select city,count(bedrooms+bathrooms), country
from airbnb_stage2
group by city,country ;
```

Step 4: Visualization

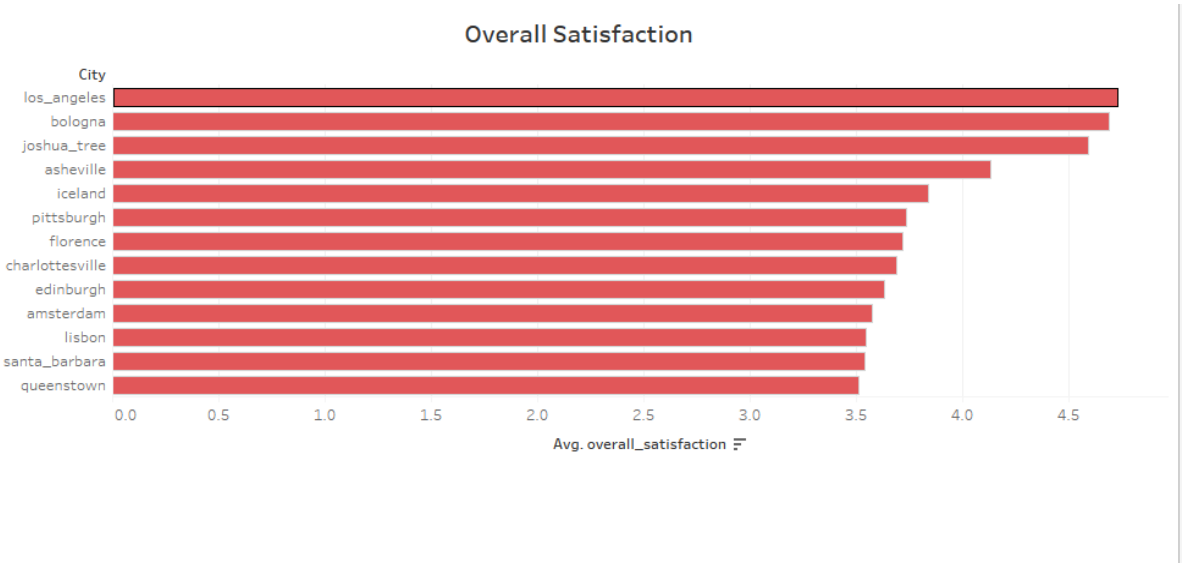
Here we are going to visualize Airbnb data using different maps and derive interesting insights which are listed below

1. Open Tableau and Upload the tables satisfaction, highaccommodates and join them. Drag drop the columns as shown in the below figure

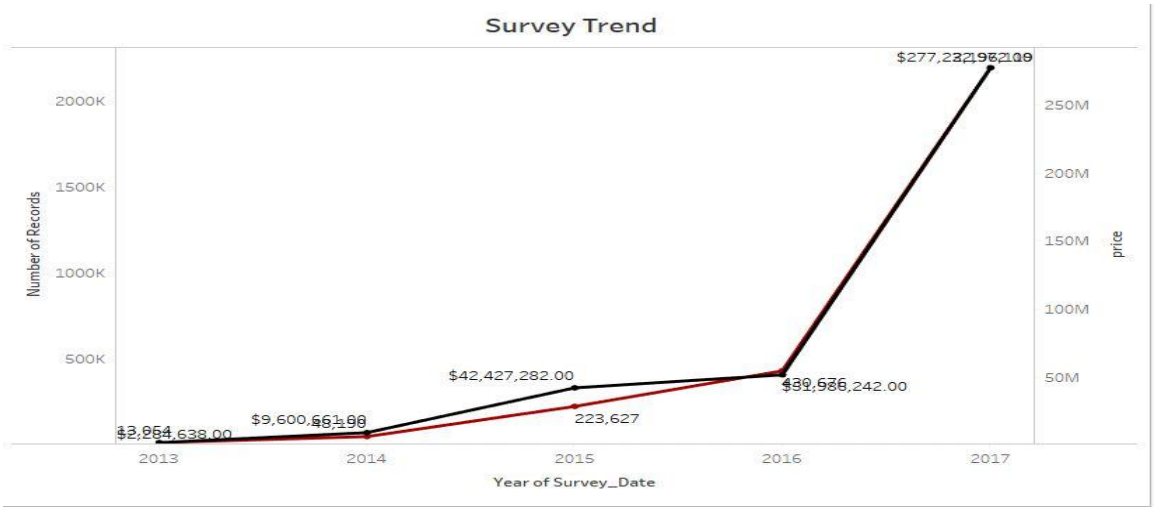
2. In this visualization, Geographic distribution with maps is used for getting insights about satisfaction of different cities around the world. Here cities are categorized on basis of least, moderate and most satisfaction.



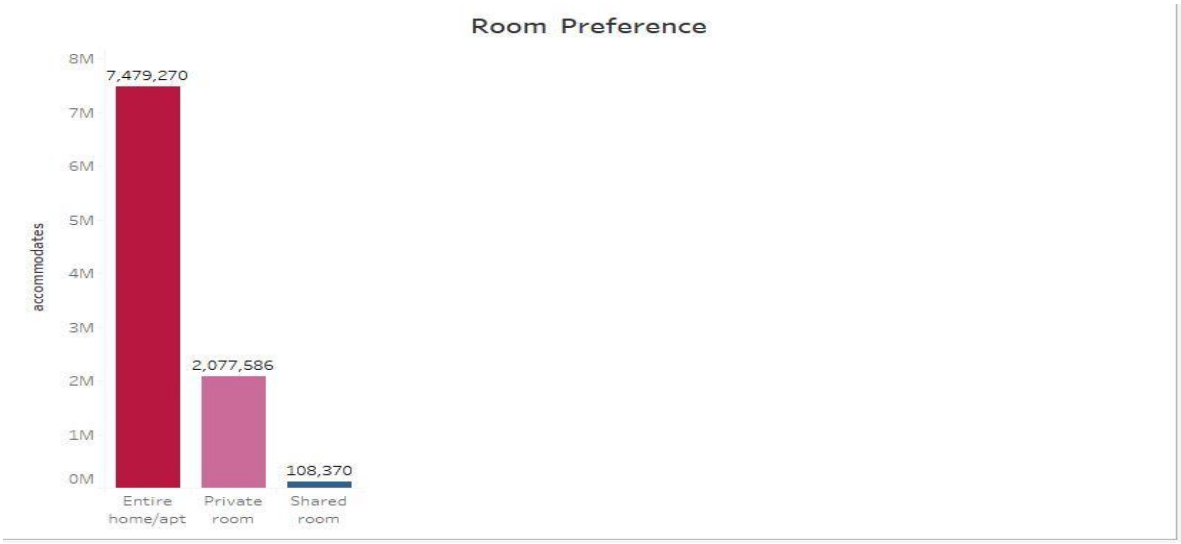
3. This map help to focus on cities with most satisfaction to look for factors of positivity and apply those factors to least satisfied cities. This is important information which will help to make better decisions.



4. Open tableau and upload the review table, rooms table and join them. Drag drop the columns as shown in the figure below
5. This Trend shows Price, which is being offered by people to the services offered. 2016 shows fall in the trend but in 2017 it again rises, as per the google study 100 million people registered for the Airbnb in the year 2017.



6. This map is going to give detailed explanation about room preferences of different cities around the world



References

1. URL of Data Source: **http://tomslee.net/airbnb-data-collection-get-the-data**
2. URL of your Github: **https://github.com/AlekhyaRaidu/CIS-5200**
3. URL of References

   Guide for Hive: **https://box/9zduux2so6z2xk461euzklmupck7vix2**

   Tableau Guide: **https://app.box.com/s/rz05haaqyh6hhczqwe0q64ti38nw9yeb**

   Guide for Hadoop: **https://app.box.com/s/5nuez5z9b9zetufmngmllf0kpff4j2yx**