IDS 561 Final Project Report

Exploring Airbnb Data of major US cities

Anuj Chanchlani - 674153696

Pratik Talreja - 657488876

Rashi Desai - 663553314

Department of Information & Decision Sciences, University of Illinois at Chicago

IDS 561: Analytics for Big Data

Prof. Yuheng Hu

December 9, 2020

PROJECT REPORT

Problem Setting

Airbnb (founded: August 2008) is a vacation rental online marketplace that offers arrangements for lodging, primarily homestays, or tourism experiences. For a company with more than 150 million users and 400 million guests since launch, data analysis is quintessential.

Airbnb is an online marketplace for people who are looking for accommodations. It connects travelers with Airbnb hosts who want to rent out their homes or other property. Airbnb, a $33.8 billion industry in the United States alone inspired our team to explore the Airbnb data and listings for major US cities using BigData platforms. We zeroed out on a problem statement to create a model that identifies Airbnb listings with similar amenities based on preferences entered by a user.

As every problem is backed up by a solid "why", we concentrated on defining a purpose behind doing this project: To help users find listings that offer similar amenities and services in a particular neighborhood.

Data Description

Dataset: Airbnb Data (New York City, Chicago, Boston, Denver)

Type: Public

Data source: Inside Airbnb (http://insideairbnb.com/get-the-data.html)

Description about data:

1. 106 attributes with 68,271 unique listings
2. A mix of categorical, object and numerical data variables
3. The data has listings from four cities (NYC, Chicago, Boston, Denver) as of August 2019
4. Listing types: apartment, house, town house, condominium
5. Important descriptive attributes:

- Amenities: WiFi, stove, hot water, bed linens)
- Room type (private, shared)
- Host verification (email, phone, reviews, offline government ID)
- Description of the listings among others

Techniques

- Coding environment: Python3 and PySpark
- Python Libraries: Pandas, Numpy, Matplotlib, Sklearn.
- Big Data tools and libraries: Apache Spark, MLlib ,Plotly.

As for any data projects, there is a life cycle followed from gathering the data, all the way up to the analysis and presenting the results, we followed the OSEMN framework - Obtain data, scrub, explore, model & interpret data for our project.

- **Data Procurement**

  We started with the first obvious step of data collection for a few of the major US cities popular with Airbnb: NYC, Chicago, Boston and Denver. Data collected for the four cities was concatenated to form a comprehensive dataset.

- **Data Cleaning**

  For preparing the data, we needed to detect and correct corrupt or inaccurate records from the combined dataset. We referred to identifying incomplete, incorrect, inaccurate or irrelevant components of the data and cleaned coarse data as:

  - Removed punctuations ($ price, comma, dots, unwanted symbols)
  - Data type conversion (string to float)
  - Removed attributes with null values >20%

- **Data Exploration**

  As part of the exploratory data analysis, we performed first-hand analysis on the Airbnb data for all four cities: New York, Chicago, Boston, Denver.

Next, we performed descriptive statistics such as:

- Finding mean price of listings using barplots and other visualizations
- Cursory data analysis for all the four cities, and for the next steps, we honed more on New York City for cluster analysis.

- **Feature selection**

  As part of this step, we selected a subset of the original features and removed unimportant variables by observation. Attributes as listing_URL, scrape_ID, country code, first & last reviews

- **Feature engineering**

  To improve the performance of our machine learning algorithm, we extracted features from raw data via data modelling. The primary task in feature engineering were:

  - One-hot encoding of categorical variables: the variables were converted into a form that could be provided to ML algorithms to do a better job in prediction and
  - Standardizing numerical variables: rescaled numeric values of original data attributes to have equal range and/or variance
  - Principal Component Analysis: To reduce complexity of a model and avoid overfitting, we moved with dimensionality reduction using PCA. With PCA, we worked to map to data to maximum the variance

- **Cluster analysis**

  We used K-means clustering to cluster the Airbnb listings based on price of listings as low, medium, high. As we completed our model build, we concluded that choosing another member of the same cluster closer to a neighborhood, say Manhattan, you can trust that there is some similarity between these two locations. The algorithm can then narrow down your search and find what a user is looking for.

The clusters predicted post our data analysis on Airbnb data takes into account descriptive attributes and provides listings that best fits a user's preference about the fields we discussed.

Results

In general data analysis of New york, Chicago, Boston, Denver listings.

a) Top 5 popular localities with most number of listings

count	city
21727	New York
19012	Brooklyn
8620	Chicago
5996	Boston
4481	Denver

b) Top 5 popular property types

property_type	count
Apartment	48595
House	7655
Condominium	3834
Townhouse	2630
Loft	1799

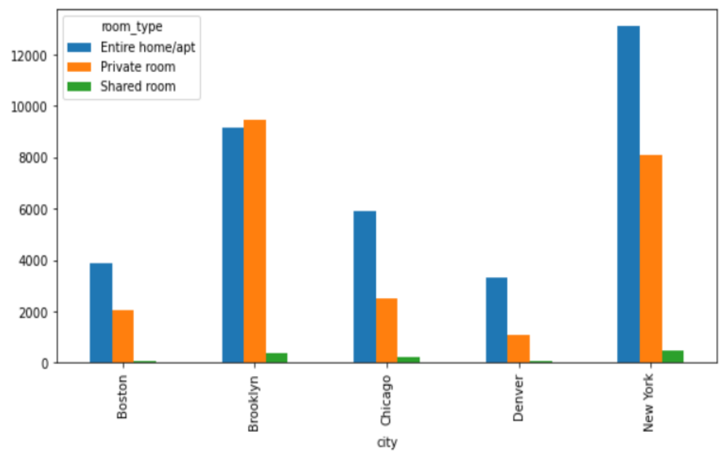
c) Distribution of these property types in top 5 localities

property_type	count	city
Apartment	19122	New York
Apartment	14682	Brooklyn
Apartment	5074	Chicago
Apartment	3970	Boston
House	1701	Denver
House	1564	Brooklyn
Condominium	1305	Chicago
House	1150	Chicago
Townhouse	923	Brooklyn

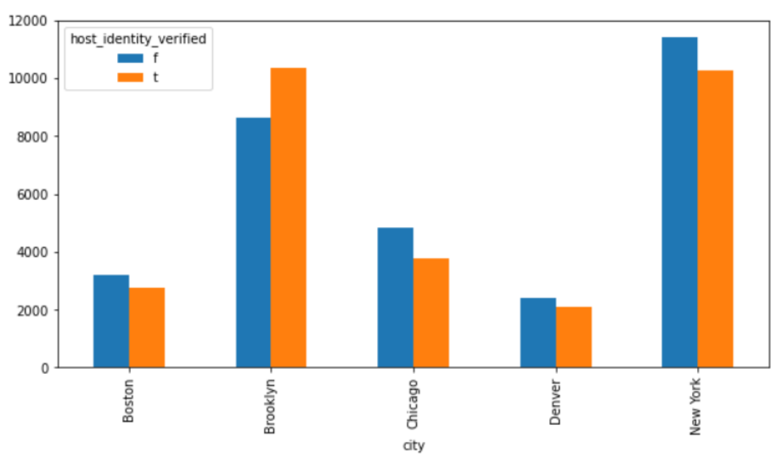
d) Top 3 popular room types

room_type	instances
Entire home/apt	38552
Private room	28186
Shared room	1532

e) Distribution of room types in top 5 localities



f) Counting number of verified and not verified hosts in top 5 localities



In the top 5 localities, except Brooklyn, the number of hosts whose identity is not verified are more.

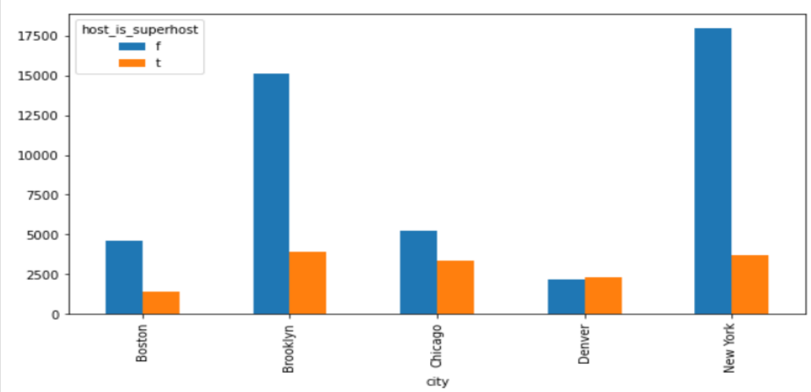
g) Counting total number of hosts with and without profile pic in top localities.

count	city	host_has_profile_pic
21656	New York	t
18985	Brooklyn	t
8613	Chicago	t
5984	Boston	t
4471	Denver	t
62	New York	f
19	Brooklyn	f
11	Boston	f
7	Chicago	f
5	Denver	f

As we can see, even though the host identity is not verified for the majority of the listings, almost all of them have the hosts profile picture. Very few hosts in major localities don’t have a profile pic.

h) Counting super hosts and not super hosts in top 5 localities

```
<matplotlib.axes._subplots.AxesSubplot at 0x220203a4860>
```



We can conclude: A host being a superhost or not is NOT what potential renters are looking for.

i) Let's check whether number of reviews of listings in top5 localities has any role to play

frequency	city
11438	New York
8650	Brooklyn
4824	Chicago
3221	Boston
2393	Denver

frequency	city
10354	Brooklyn
10280	New York
3796	Chicago
2774	Boston
2083	Denver

From the tables we can say that, the listings whose host have no identity verified when grouped by along top5 cities, have received sufficient reviews. So we can say that along with a profile picture of the host, the number of reviews a listing has received has some role to play for it to get booked.

j) Counting number of listings in each review score band for both host identity = true and false

review_scores_value	frequency
10.0	14225
9.0	8437
8.0	1430
7.0	196
6.0	172
4.0	39
2.0	37
5.0	16
3.0	4

review_scores_value	frequency
10.0	12694
9.0	7929
8.0	1642
7.0	272
6.0	262
2.0	79
4.0	65
5.0	24
3.0	3

Both types of listings (with host verified and not verified) have good review scores hence strengthening our assumption that review scores play an important role for bookings.

k) Now, let's see if pricing of listings across top 5 cities plays an important role in booking.

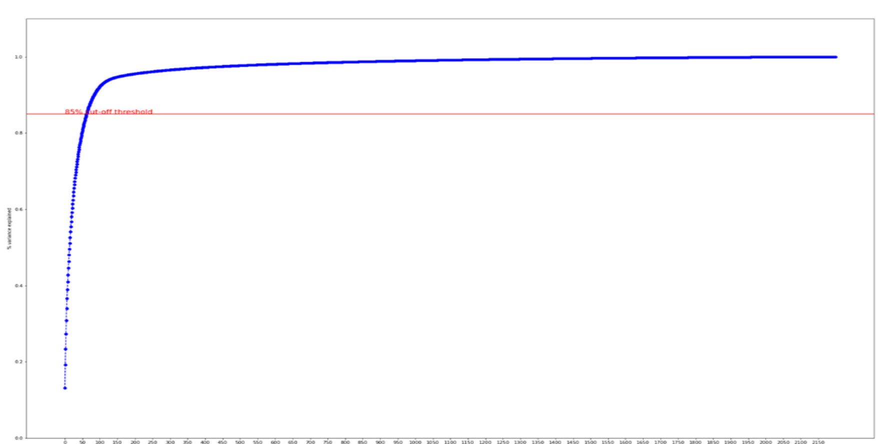
city	average_price
New York	199.41991606924287
Boston	195.78298665010865
Denver	188.66485582950273
Chicago	188.48175787728027
Brooklyn	122.50820809248555

city	average_price
Boston	207.1240086517664
New York	186.9681906614786
Chicago	161.8427291886196
Denver	132.8900624099856
Brooklyn	126.178868070311

If we compare averages prices of listings across top 5 localities with host_identity_verified = false and host_identity_verified = true, we can see that even though listings whose host_identity is not verified, still have higher prices. But sufficient reviews, good review scores and hosts having profile pictures might be sufficient to get a decent number of bookings.

Cluster analysis on New York specific listings

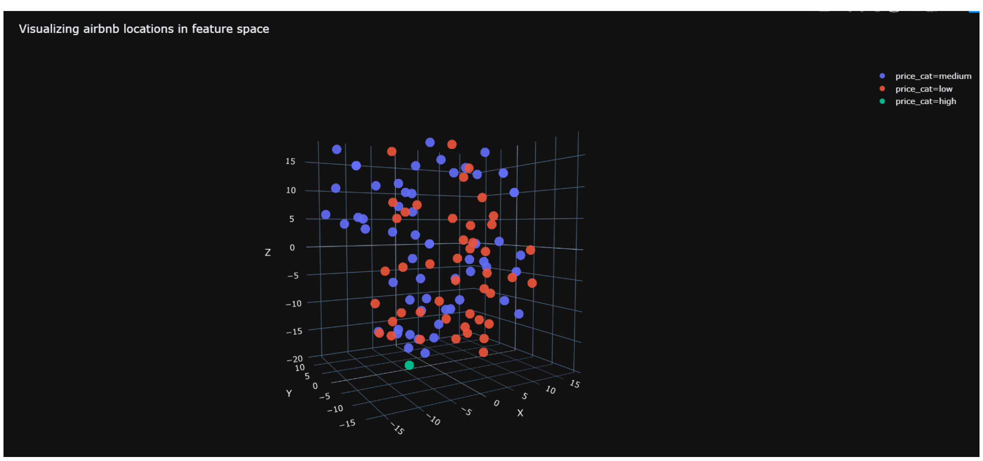
a) Selecting number of of components for PCA *(**Note:** y axis % variance , x axis number of components)*



b) Principal Component Analysis

```
+--------------------+--------------------+
|            features|         pcaFeatures|
+--------------------+--------------------+
|(2200,[0,1,2,5,6,...|[-2.2815516044260...|
|(2200,[0,1,2,3,4,...|[-0.1029831706585...|
|(2200,[0,1,2,3,5,...|[-4.2849699572688...|
|(2200,[0,1,2,3,5,...|[-1.8187386209920...|
|(2200,[0,1,2,3,5,...|[-4.3754383224395...|
|(2200,[0,1,2,3,4,...|[-0.2050388367894...|
|(2200,[0,1,3,4,5,...|[-1.1846446797933...|
|(2200,[1,2,3,4,5,...|[-0.5823883083608...|
|(2200,[0,1,2,3,5,...|[-1.4773030840515...|
|(2200,[0,1,2,3,9,...|[-1.3053351830624...|
|(2200,[0,1,2,3,4,...|[-4.4384163871801...|
|(2200,[0,1,2,3,5,...|[-3.1488054282759...|
|(2200,[0,1,3,7,8,...|[-1.1570501995614...|
|(2200,[0,1,2,3,5,...|[-5.3090319091476...|
|(2200,[0,1,2,3,4,...|[-2.6598761092262...|
|(2200,[0,1,2,3,12...|[-0.2714055255642...|
|(2200,[0,1,2,3,5,...|[-2.4571749819526...|
|(2200,[0,1,2,3,4,...|[-2.1352912338178...|
|(2200,[0,1,2,3,6,...|[-0.4208247418201...|
|(2200,[1,2,3,8,9,...|[-0.5249989755709...|
+--------------------+--------------------+
```

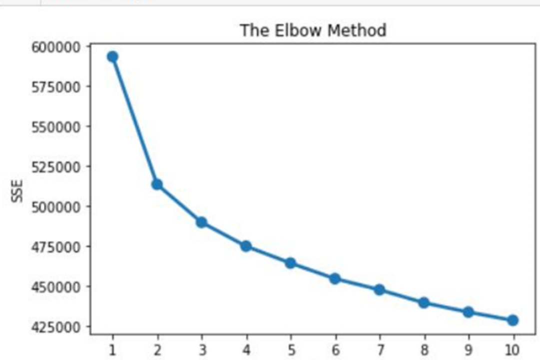
c) Visualizing listings in 3D space *(**Note:** .html file of above visualization is uploaded along with the code)*



d) Elbow Plot : K-means

n [20]:

```
plt.title('The Elbow Method')
plt.xlabel('k')
plt.ylabel('SSE')
sns.pointplot(x=list(sse.keys()), y=list(sse.values()))
plt.show()
```



e) Clustering

```
predictions.show(100)
```

```
+--------------------+--------------------+----------+
|            features|         pcaFeatures|prediction|
+--------------------+--------------------+----------+
|(2200,[0,1,2,5,6,...|[-2.2815516044260...|         2|
|(2200,[0,1,2,3,4,...|[-0.1029831706585...|         0|
|(2200,[0,1,2,3,5,...|[-4.2849699572688...|         1|
|(2200,[0,1,2,3,5,...|[-1.8187386209920...|         2|
|(2200,[0,1,2,3,5,...|[-4.3754383224395...|         1|
|(2200,[0,1,2,3,4,...|[-0.2050388367894...|         0|
|(2200,[0,1,3,4,5,...|[-1.1846446797933...|         2|
|(2200,[1,2,3,4,5,...|[-0.5823883083608...|         0|
|(2200,[0,1,2,3,5,...|[-1.4773030840515...|         2|
|(2200,[0,1,2,3,9,...|[-1.3053351830624...|         2|
|(2200,[0,1,2,3,4,...|[-4.4384163871801...|         1|
|(2200,[0,1,2,3,5,...|[-3.1488054282759...|         2|
|(2200,[0,1,3,7,8,...|[-1.1570501995614...|         0|
|(2200,[0,1,2,3,5,...|[-5.3090319091476...|         1|
|(2200,[0,1,2,3,4,...|[-2.6598761092262...|         2|
|(2200,[0,1,2,3,12...|[-0.2714055255642...|         0|
|(2200,[0,1,2,3,5,...|[-2.4571749819526...|         1|
```

Role of Team Members in the Project

Team Member Name	Role of member
Anuj Chanchlani	Feature engineering and cluster analysis, Outcome visualization Log project issues and create contingency plans
Pratik Talreja	Research data sources, procure data and data quality check Exploratory Data Analysis , Maintained team documentation
Rashi Desai	Data preparation (cleaning, granularity, validation), first-hand data modelling and feature selection, tracked project milestones