Capstone Project - The Battle of Neighborhoods

Deciding The Neighborhood Due To Relocation And Finalizing Amenities For Settlement With Family

Abha Vasal

1. Introduction

1.1. Background

Shifting to a new city is never an easy decision. We analyze the cost of living parameter before moving in. Relocation to a different city means finalizing the neighborhood that is safe for living and provides convenient access to basic amenities. There is an immediate need for accommodation and children must be admitted to schools. Finalizing accommodations takes time and once has to look for temporary accommodation at a Hotel or Airbnb complexes.

1.2. Problem Definition

At times we are in a dilemma whether to relocate to a city on getting a promotion or not. With relocation we have to look for accommodations in safer neighborhoods for our family and schools and other amenities for our family.

As my problem statement I explore the case of person working for XYZ Corp. On getting a promotion he has the option of relocating to either Boston or Chicago, as his team works from those locations. Shifting from Dallas, he must analyze which is the best option for him in terms of cost of living index, property rent index and affordability ratio etc.

With his family moving in with him he needs to look for Residential Complexes for settling down. We will analyze the crime record for the year 2018 and look for safer neighborhoods in Chicago. We will explore the neighborhoods and look for areas with amenities like park, convenience stores, Bus station, Train station nearby. The project aims at suggesting neighborhoods for staying. With his children studying in Elementary and High School. He needs to find schools in his neighborhood.

The problem can recommend other people thinking of relocating and suggest them ways to explore the neighborhood and eventually settling down.

Once the city is finalized and before moving into final Residential Complex, for temporarily accommodation we aim at suggesting different Airbnb options available. Through parameters like neighborhood, property type, different amenities, the project will try aiming at predicting price of accommodations.

2. Describe the data
2.1.Data Sources

1. In this section, I will describe the data used to solve the problem as described previously.
   The website https://www.numbeo.com/ was scrapped to extract data about of cost of living index and property price index for all the major world cities. Rows corresponding to Boston and Chicago was filtered and analyzed visually. Using BeautifulSoup and Requests, the results the results were retrieved.

2. Chicago is divided into 77 community areas. Community areas are distinct from the more numerous neighborhoods in Chicago. Community areas often encompass groups of neighborhoods. Although many community areas contain more than one neighborhood, they may also share the same name, or parts of the name, of some of their individual neighborhoods.
   Wikipedia site https://en.wikipedia.org/wiki/Community_areas_in_Chicago was used to extract names of Community Areas and neighborhoods with in those areas.
3. We used open source Chicago Crime data from https://data.cityofchicago.org to provide the user with crime data for the year 2018.

   A record looks like this

   ID 11561837
   Case Number JC110056
   Date 12/31/2018 11:59:00 PM
   Block 013XX W 72ND ST
   IUCR 1153
   Primary Type DECEPTIVE PRACTICE
   Description FINANCIAL IDENTITY THEFT OVER $ 300
   Location Description
   Arrest
   Domestic
   Beat 0734
   District 007
   Ward 6
   Community Area 67
   FBI Code 11
   X Coordinate 1168573
   Y Coordinate 1857018
   Year 2018
   Updated On 01/17/2019 02:26:36 PM
   Latitude 41.763181359
   Longitude-87.657709477
   Location

4. We will Query the FourSqaure website www.foursquare.com to explore the neighborhood of Chicago Use the FourSquare API to look for Residential Complexes in finalized neighborhoods
5. Use the FourSquare API to search for schools in nearby neighborhood

   The project aims at suggesting which city to shift to in terms of better quality of life and having less crime rate.

   The study will help in identifying and recommending neighborhood seeing the needs of his family and crime rate in that neighborhood using the Foursquare API.
   For immediate relocation the project will also recommend Airbnb accommodations available in those neighborhood
   url='https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&ll={},{}&radius={}&limit={}'.format(CLIENT_ID, CLIENT_SECRET, VERSION, neighborhood_latitude, neighborhood_longitude, radius, LIMIT)

   Send the GET request and examine the results
   results = requests.get(url).json()

6. Data about Chicago Airbnb listings data is downloaded from http://insideairbnb.com/get-the-data.html. The data behind the Inside Airbnb site is sourced from publicly available information from the Airbnb site. The file listings – provides detailed listings showing 96 attributes for each of the listings.
   Some of the attributes used in the analysis are price (continuous), longitude), latitude (continuous), property_type (categorical), is_superhost (categorical), neighborhood , ratings (continuous) ,reviews are some of them.

2.2 Data Cleaning

Data was downloaded or scraped from multiple sources like www.numbeo.com and wikipedia. Wikipedia data was spread across multiple tables, it had to be scraped and loaded it dataframe. Crime data from https://data.cityofchicago.org/ had a lot of missing values.
The data consisted of crime records with respect to wards, districts and community areas. Chicago is divided into 77 community areas. Community areas often encompass groups of neighborhoods. To identity the neighborhoods the data scraped from Wikipedia was to identify the community areas with Chicago neighborhoods.

CommunityCode	CommunityArea	Neighbourhood
08	Near North Side	Cabrini–Green\nThe Gold Coast\nGoose Island\nM...
32	Loop	Loop\nNew Eastside\nSouth Loop\nWest Loop Gate
33	Near South Side	Dearborn Park\nPrinter's Row\nSouth Loop\nPrai...
05	North Center	Horner Park\nRoscoe Village
06	Lake View	Boystown\nLake View East\nGraceland West\nSout...

The neighborhood data collected by separated by '\n' , it was replaced by commas and finally the neighborhoods were split in multiple rows.

Merging Dataframes to include Community Code, Community area along with details of Neighborhood

```
finalDataSet = pd.DataFrame.merge(ca_df, fig2d,how='inner', left_on='CommunityCode',right_on='Community_Area')
finalDataSet.head()
```

	CommunityCode	CommunityArea	Neighbourhood	Community_Area	Count_per_Community_Area
0	8	Near North Side	Cabrini–Green,The Gold Coast,Goose Island,Magn...	8	13079
1	32	Loop	Loop,New Eastside,South Loop,West Loop Gate	32	10879
2	33	Near South Side	Dearborn Park,Printer's Row,South Loop,Prairie...	33	1876
3	5	North Center	Horner Park,Roscoe Village	5	1331
4	6	Lake View	Boystown,Lake View East,Graceland West,South E...	6	5961

```
finalDataSet= finalDataSet.sort_values(by='Count_per_Community_Area', ascending=True).reset_index(drop=True)
finalDataSet.head()
```

	CommunityCode	CommunityArea	Neighbourhood	Community_Area	Count_per_Community_Area
0	9	Edison Park	Edison Park	9	253
1	47	Burnside	Burnside	47	386
2	12	Forest Glen	Edgebrook,Old Edgebrook,South Edgebrook,Saugan...	12	496
3	74	Mount Greenwood	Mount Greenwood Heights,Talley's Corner	74	570
4	18	Montclare	Montclare	18	603

Like wise Airbnb data was suggesting accommodations in regard to neighborhoods so Wikipedia data was helpful in interpreting.

Crime records had missing values. Values of XCoordinate, YCoordinate, latitudes , longitudes and location were missing for many records, they had to be removed for plotting choropleth maps. Location column was duplicate information as longitude latitude were mentioned separately.

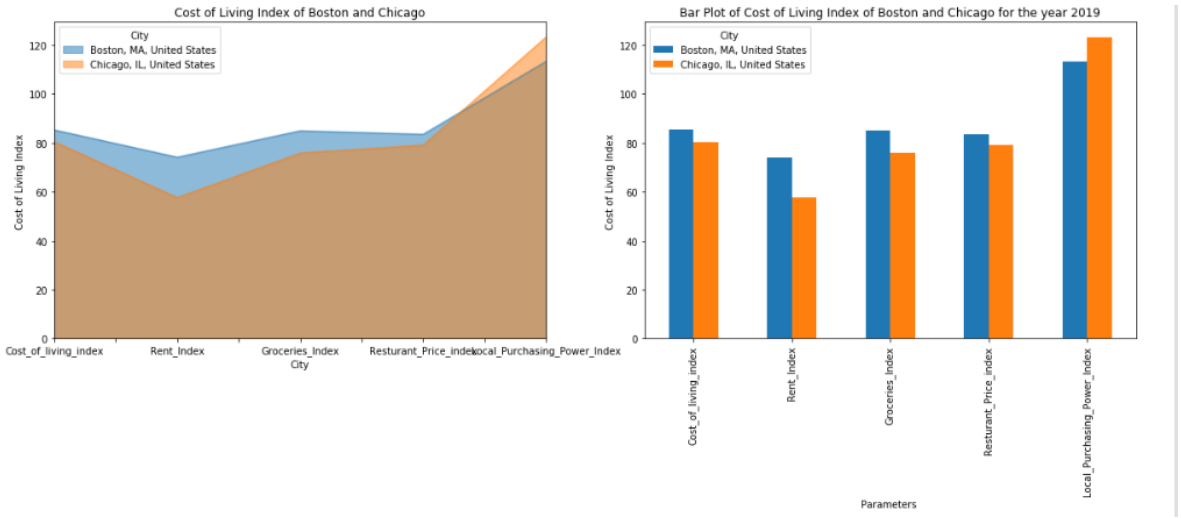
Airbnb datasets also had missing values in many columns like bathrooms,bedrooms, review_scores_rating, cleaning fee, security_deposit, host_location, host_abouthost_response_time. Data was analyzed to check if these rows can be ignored or should be filled with zero or median or mean values. Slightly correlated features were kept, others were dropped from the dataset.

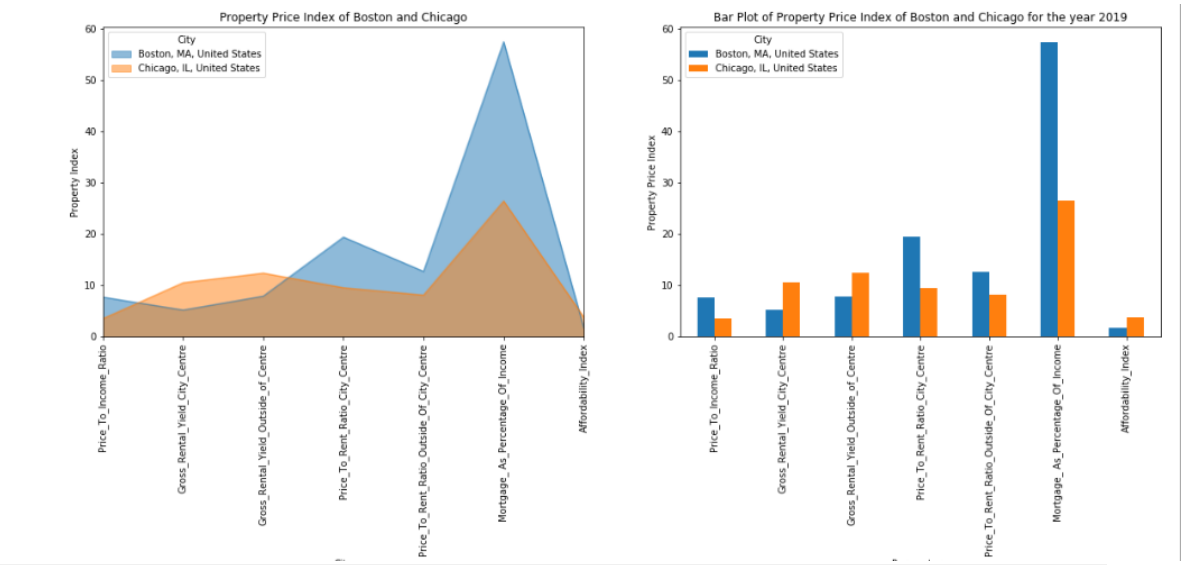
3. Exploratory Data Analysis

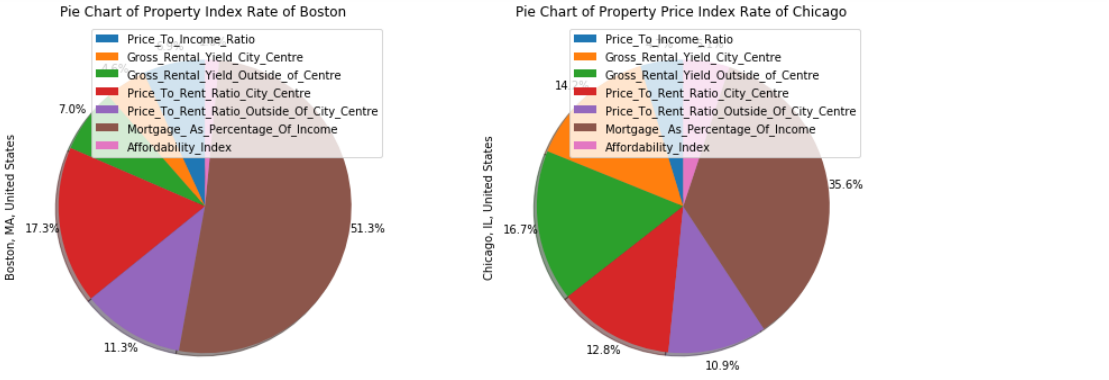
3.1. Determining relocation city

Data from numbeo website was extracted and loaded into a dataframe. Visually compared with the help of bar plot ,area plot and pie charts to show distribution.

City	Boston, MA, United States	Chicago, IL, United States
Price_To_Income_Ratio	7.69	3.51
Gross_Rental_Yield_City_Centre	5.16	10.50
Gross_Rental_Yield_Outside_of_Centre	7.87	12.36
Price_To_Rent_Ratio_City_Centre	19.38	9.52
Price_To_Rent_Ratio_Outside_Of_City_Centre	12.70	8.09
Mortgage_ As_Percentage_Of_Income	57.44	26.41
Affordability_Index	1.74	3.79







We can obeserve cost of living in Chicago is lower. Purchasing Power is higher in Chicago city as compared to Boston

Cost of Property Prices, rents are higher in Boston.

3.2. Determining Safe community areas

The crime records consisted of incidents occurring in regard to district, ward and community area. To find safer community areas hence safe neighborhoods. The data had to be grouped district wise and community area number wise.

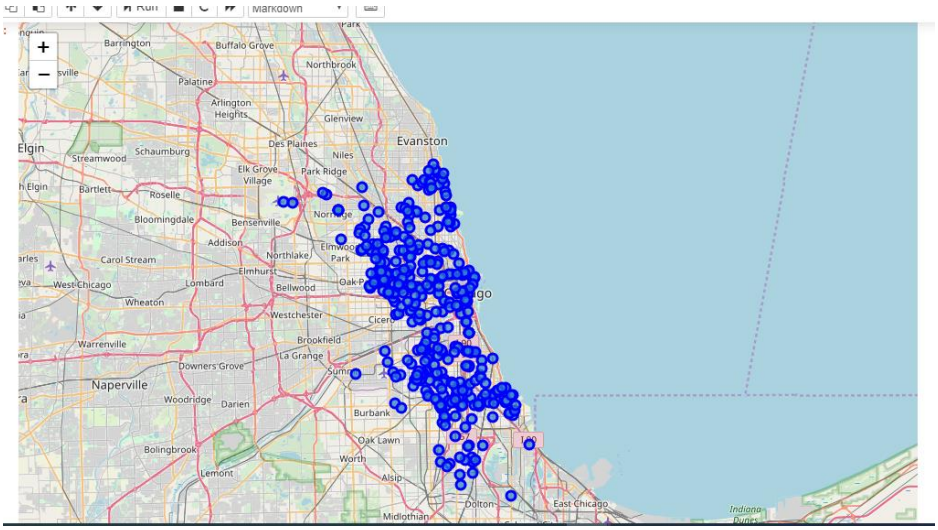


Figure showing folium map of crime records of Chicago

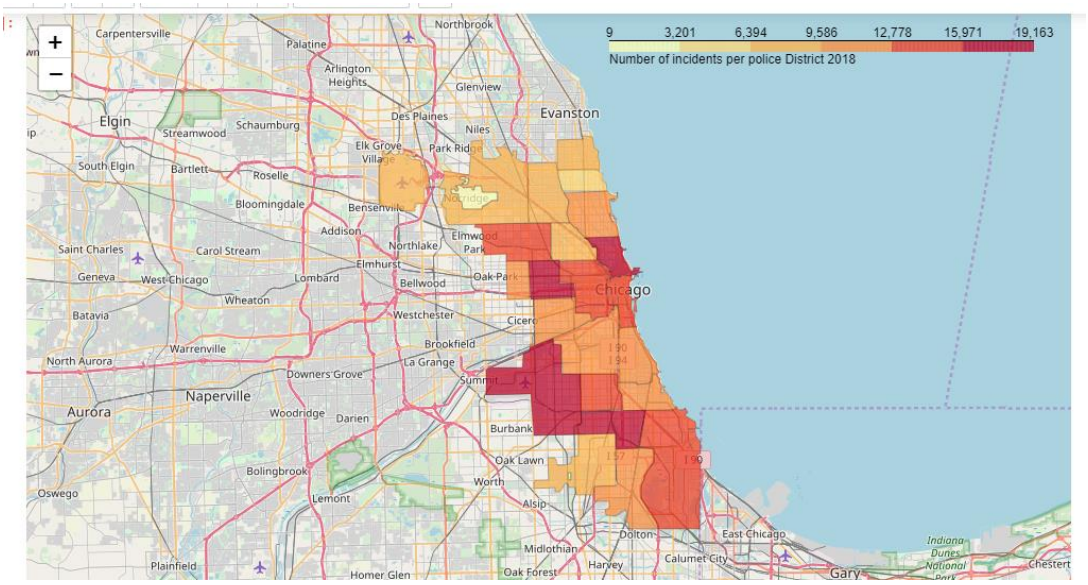
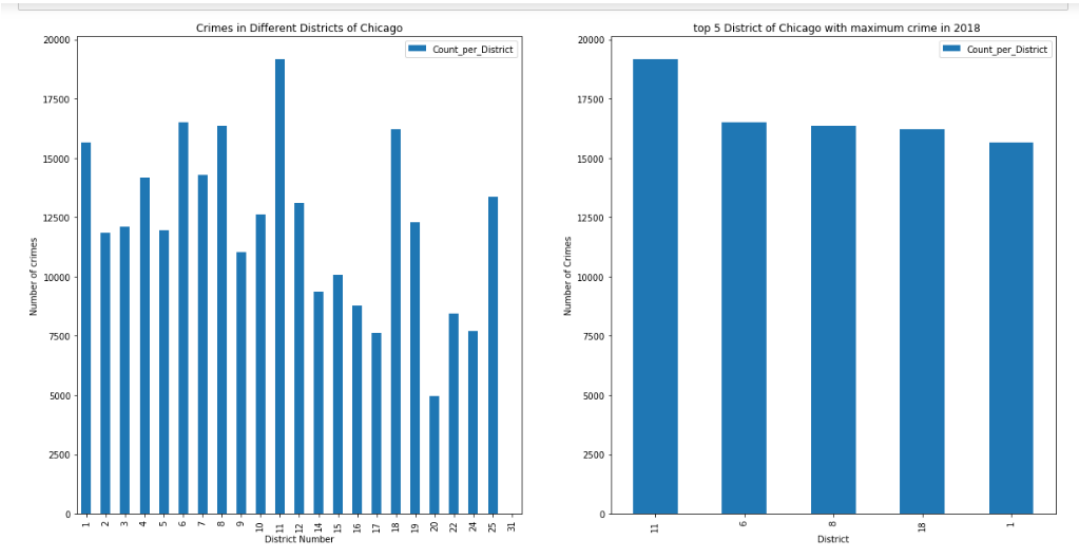
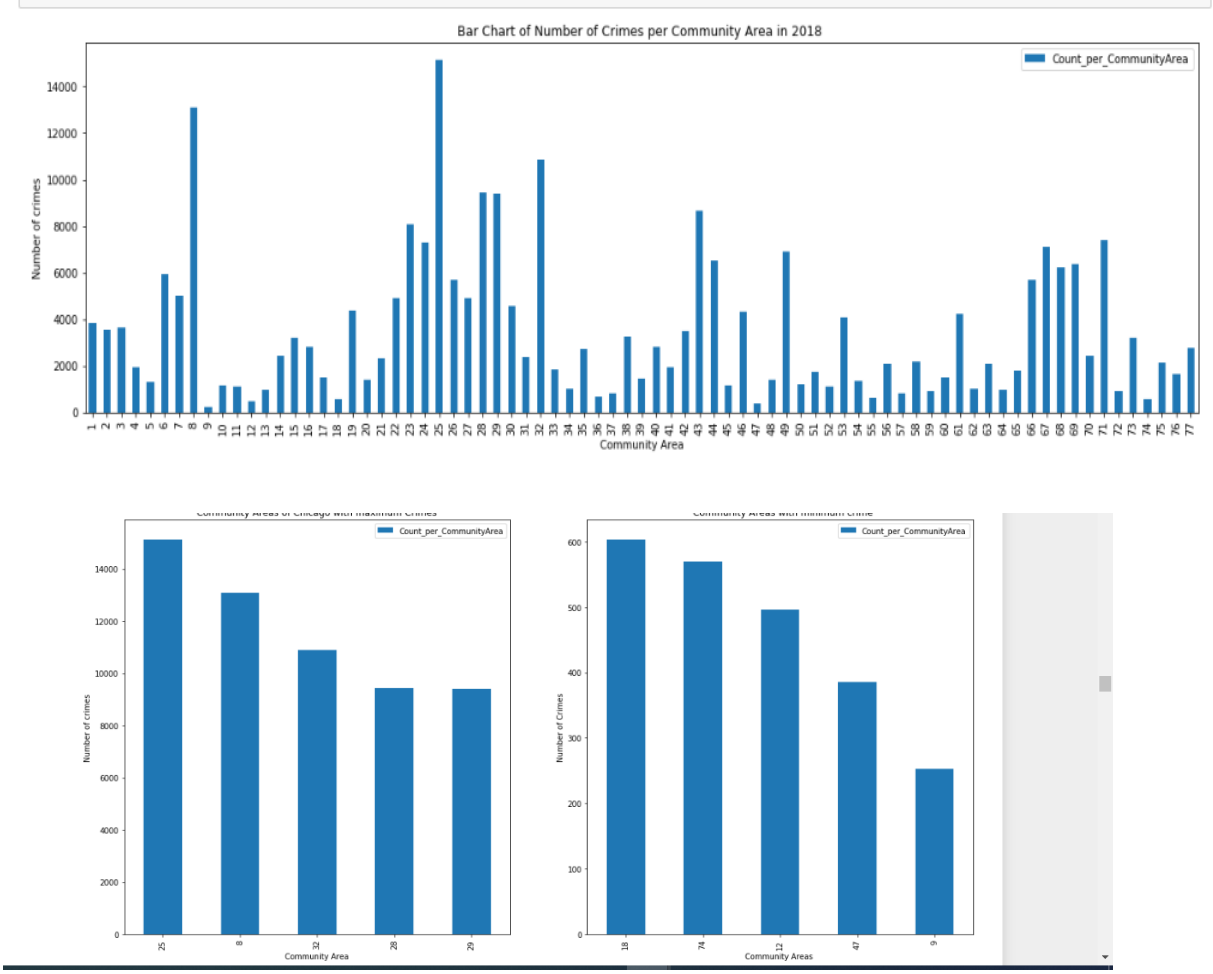


Figure representing number of crimes per police district



<Figure size 432x288 with 0 Axes>

It was observed that districts 11, 6, 8, 18 and 1, had maximum number of crimes. Community wise grouping was done to identify neighborhoods with maximum and minimum crimes.



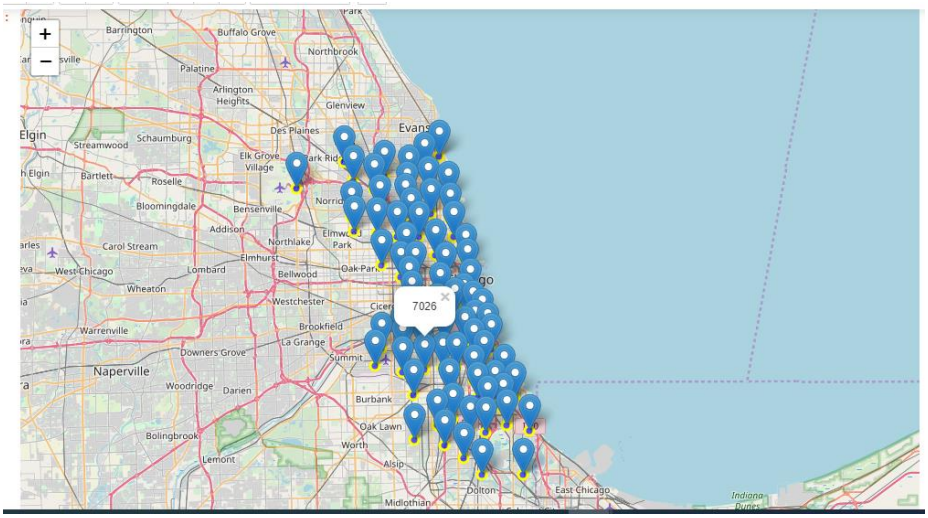


Figure showing crime in different community areas. Markers showing the count in the area

The top 5 safest community Areas were 9, 47, 12, 74 and 18. Community Areas Edison Park, Burnside, Forest Glen, Mount Greenwood, Montclare are safer areas.

3.3. Exploring the Neighborhoods

Merging crime data and neighborhood data, coordinates for community areas were determined. **S**afest community areas data, was used to explore neighborhoods and search nearby venues using the Foursquare API. The venues were searched with in the radius of 500m. The get_category_type() function was used for categorising the venues. There are 71 unique categories.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Archer Heights	Mexican Restaurant	Bakery	Gas Station	Hot Dog Joint	Pharmacy	Nightclub	Cosmetics Shop	Mobile Phone Shop	Rental Car Location	Rental Service
1	Beverly	Park	Souvenir Shop	Wings Joint	Discount Store	Dry Cleaner	Fast Food Restaurant	Financial or Legal Service	French Restaurant	Fried Chicken Joint	Gas Station
2	Brynford Park	Korean Restaurant	Sandwich Place	Bus Station	Coffee Shop	Park	Japanese Restaurant	Radio Station	College Quad	College Bookstore	Soccer Field
3	Burnside	Intersection	Train Station	Convenience Store	Construction & Landscaping	Wings Joint	Gas Station	Fast Food Restaurant	Financial or Legal Service	French Restaurant	Fried Chicken Joint
4	Chrysler Village	Video Store	Pizza Place	American Restaurant	Convenience Store	Gay Bar	Fried Chicken Joint	Mexican Restaurant	Nightclub	Fast Food Restaurant	Donut Shop

Each neighborhood was analysed. Rows were grouped by neighborhood and then the mean of the frequency of occurrence of each category was determined.

	Neighborhood	American Restaurant	Asian Restaurant	Bakery	Bar	Baseball Field	Boutique	Breakfast Spot	Bus Station	Chinese Restaurant	Coffee Shop	College Bookstore	College Quad	Construct Landscap
0	Archer Heights	0.000000	0.000000	0.142857	0.047619	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0
1	Beverly	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0
2	Brynford Park	0.000000	0.041667	0.041667	0.041667	0.000000	0.000000	0.041667	0.083333	0.000000	0.083333	0.041667	0.041667	0
3	Burnside	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0
4	Chrysler Village	0.055556	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.055556	0.000000	0.000000	0.000000	0
5	Edgebrook	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0
6	Edison Park	0.000000	0.000000	0.000000	0.062500	0.000000	0.000000	0.062500	0.000000	0.000000	0.000000	0.000000	0.000000	0

Top 10 common venues for each Neighborhood was determined. The venues were clustered using kmeans method in to 4 clusters. A new dataframe was create that includes all the details of the neighborhood as well as the top 10 venues for each neighborhood. Each cluster was explored. It was found that Cluster 2 has Convenience store eating joints gas station Bus station near it. Comfortable for staying

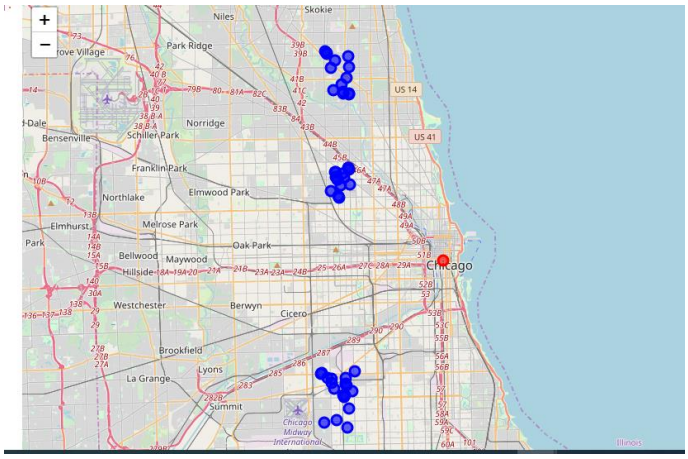
3.4. Looking for schools in neighborhoods of cluster 2

For admission of children in nearby schools. Each neighborhood of cluster 2 was explored using the Foursquare API. A function getNearbySchool() was defined to search for schools.

search_query = 'school'
radius = 1000
url='https://api.foursquare.com/v2/venues/search?client_id={}&client_secret={}&ll={},{}&v={}&query={}&radius={}&limit={}'.format(CLIENT_ID, CLIENT_SECRET, lat, lon, VERSION, search_query, radius, LIMIT)
Information of interest was determined and filtered by analyzing results which included the word school in them.

	name	categories	address	cc	city	country	crossStreet	distance	formattedAddress	labeledLatLngs	lat	lr
0	Administration Building, Lincolnwood School Di...	Government Building	6950 N East Prairie Rd	US	Lincolnwood	United States	NaN	836	[6950 N East Prairie Rd, Lincolnwood, IL 60712...	[{'label': 'display', 'lat': 42.00662783127496...	42.006628	-87.7264
1	Todd Hall, Lincolnwood School District 74	School	3925 W Lunt Ave	US	Lincolnwood	United States	NaN	958	[3925 W Lunt Ave, Lincolnwood, IL 60712, Unite...	[{'label': 'display', 'lat': 42.00733958885679...	42.007340	-87.7278
2	JCB Yeshiva - Rabbi Doug's School	Student Center	3145 W Pratt Blvd	US	Chicago	United States	Kedzie	645	[3145 W Pratt Blvd (Kedzie), Chicago, IL 60645...	[{'label': 'display', 'lat': 42.00464370325810...	42.004644	-87.7089
3	Roza's School Of Nail Technology	Trade School	NaN	US	Chicago	United States	NaN	1161	[Chicago, IL, United States]	[{'label': 'display', 'lat': 41.99742115950946...	41.997421	-87.7233
4	Rutledge High School	High School	NaN	US	Lincolnwood	United States	NaN	857	[Lincolnwood, IL, United States]	[{'label': 'display', 'lat': 42.0057308, 'lng'	42.005731	-87.7266

There were duplicates in results. Duplicate rows were ignored. The final result consisted of 66 schools. The results were visualized on the map.



3.5. Searching for Housing Complex

For finally settling down in the city they were immediately in need for moving in Residential accommodations with the family. The Four square API was used to search for residential houses with in the neighborhood of cluster 2. A function getNearbyHouses() was defined to explore the cluster for Housing complexes within the radius of 1000m. The results were analyzed and filtered to obtain information of interest that included the keyword 'Residential Building (Apartment / Condo)' in category heading. Five suggestions were made.

```
def getNearbyHouses(names, latitudes, longitudes, radius=1000):

    house_list= pd.DataFrame()
    search_query ='Housing Apartment'
    radius=1000
url='https://api.foursquare.com/v2/venues/search?client_id={}&client_secret={}&ll={},{}&v={}&query={}&radius={}&limit={}'.format(CLIENT_ID, CLIENT_SECRET, lat, lon, VERSION, search_query, radius, LIMIT)
```

name	categories	address	cc	city	country	crossStreet	distance	formattedAddress	labeledLatLngs	lat	lng	postalCode
Apartmentcito 2.0	Residential Building (Apartment / Condo)	NaN	US	Chicago	United States	NaN	857	[Chicago, IL, United States]	[{'label': 'display', 'lat': 41.93393348969538...	41.933933	-87.708475	NaN
Hispanic Housing	Residential Building (Apartment / Condo)	2806 N Sawyer Ave	US	Chicago	United States	NaN	716	[2806 N Sawyer Ave, Chicago, IL 60618, United ...	[{'label': 'display', 'lat': 41.932285, 'lng':...	41.932285	-87.708958	60618
Midpointe Apartments	Residential Building (Apartment / Condo)	4050 West 115th Street	US	Chicago	United States	NaN	1263	[4050 West 115th Street, Chicago, IL 60655, Un...	[{'label': 'display', 'lat': 41.6843237, 'lng'...	41.684324	-87.721198	60655
South Gate Apartments	Residential Building (Apartment / Condo)	4050-4064 W 115th St	US	Chicago	United States	Pulaski	1033	[4050-4064 W 115th St (Pulaski), Chicago, IL 6...	[{'label': 'display', 'lat': 41.68661804923405...	41.686618	-87.721463	60655
milwaukee apartments	Residential Building (Apartment / Condo)	3060 N Milwaukee	US	Chicago	United States	NaN	954	[3060 N Milwaukee, Chicago, IL, United States]	[{'label': 'display', 'lat': 41.93691940543633...	41.936919	-87.720593	NaN

Figure showing sample results obtained

3.6. Searching Airbnb Chicago to for immediate relocation

The website Airbnb was used to explore avenues in Chicago for the time no residential complex is available for moving in. The listings consisted with data like name, amenities, room type and property type, location and host details. There were 8852 listings in total. The categorical features were explored.

```
listings.groupby(by='neighbourhood_cleansed').count()[['id']].sort_values(by='id', ascending=False).head(10)
```

neighbourhood_cleansed	id
West Town	1127
Near North Side	852
Lake View	780
Logan Square	600
Loop	527
Near West Side	449
Lincoln Park	437
Uptown	303
Lower West Side	287
Edgewater	230

The count of number of properties listed in each neighborhood, room type, property was determined.

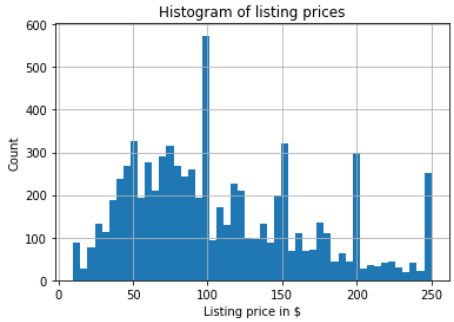
room_type	id
Entire home/apt	6022
Private room	2487
Shared room	195
Hotel room	148

```
listings.groupby(by='property_type').count()[['id']].sort_values(by='id', ascending=False).head()
```

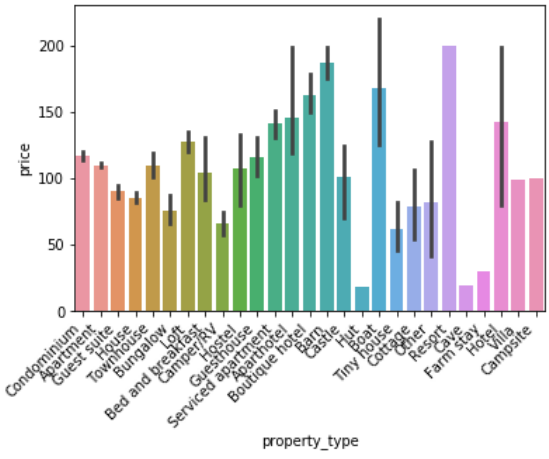
property_type	id
Apartment	5197
Condominium	1329
House	1212
Loft	244
Townhouse	234

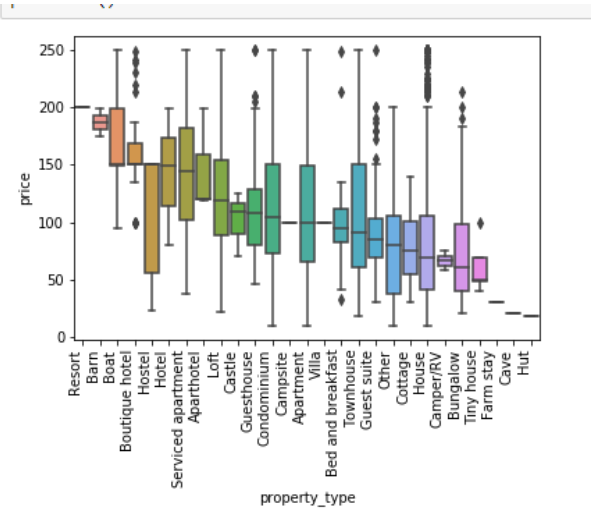
A model was create to predict the price depending on neighborhood, amenities, room type, property type etc.

The price column was explored and found that the maximum price was $ 10000/-.More than 75% of listing had price less than $200. One property listed had price zero. The outliers were removed.



Property type vs prices

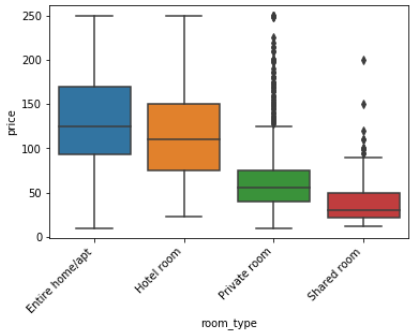




It was observed prices of condominium, apartments were less than 150$

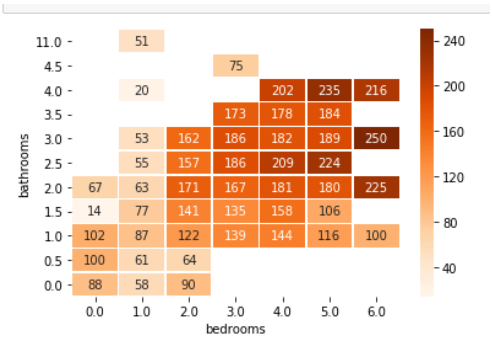
Room Type vs price

```
In [26]: ### roomtype vs price of properties having price >0 and less than 250$
sort_price = listings.loc[(listings.price <= 250) & (listings.price > 0)]\
                     .groupby('room_type')['price']\
                     .median()\
                     .sort_values(ascending=False)\
                     .index
sns.boxplot(y='price', x='room_type', data=listings.loc[(listings.price <= 250) & (listings.price > 0)], order=sort_price)
ax = plt.gca()
ax.set_xticklabels(ax.get_xticklabels(), rotation=45, ha='right')
plt.show()
```



It was observe that the average prices of entire home/apartment were highest. There were many outliers in the category private room and shared room.

Relationship between price and number of bedrooms and bathrooms



Major listings have bedrooms > 2 and bedrooms >=1

Heatmaps were used to determine the correlation between the attributes. It was seen that they have some impact on the pricing. Major listings had more than or equal to 2 bedrooms.

Numerical features were explored for missing values. Categorical variables were handled using one hot encoding. The final listing had 318 features.

4. Building predictive models

There are two types of models, regression and classification, Regression models can provide information like predicting the price given a set of features.

The final data set was used for predicting. The final 6586 listing were used for building the model. The list of few chosen features is given below.

features = list_final[['host_is_superhost', 'host_identity_verified', 'host_has_profile_pic','is_location_exact', 'requires_license', 'instant_bookable', 'require_guest_profile_picture', 'require_guest_phone_verification', 'security_deposit', 'cleaning_fee', 'host_listings_count', 'host_total_listings_count', 'minimum_nights', 'bathrooms', 'bedrooms', 'guests_included', 'number_of_reviews','review_scores_rating','price']]

I applied linear models like (linear regression, Ridge regression, random forest models to the dataset, using root mean squared error (RMSE)

a)Using Linear Model

A Linear Regression model was constructed. The listings were split into training and testing sets. 20% of samples were used for testing.

```
#Score/Accuracy
print("Accuracy --> ", model.score(X_test, y_test)*100)
print(model.score(X,y))
```

```
Accuracy -->  54.10728360510062
0.5836944917072064
```

```
print("Predicted values:", yhat_test[0:10])
print("True values:", y_test[0:10].values)
```

```
Predicted values: [ 47.92438675 147.31123031 167.98022467  68.73305221 135.10554389
  55.90738175 120.7101612  117.60973859  93.11306793  71.80663669]
True values: [ 50. 149. 183.  85. 175.  50. 180. 100.  40.  95.]
```

```
# RMSE
print(np.sqrt(metrics.mean_squared_error(y_test, yhat_test)))
```

```
39.52743782135698
```

b) Random Forrest Regression-

The following parameters were set- n_estimators=500, criterion='mse', test size was 10%

```
from sklearn.metrics import mean_squared_error
from sklearn.ensemble import RandomForestRegressor
```

```
Xtrain, Xtest, ytrain, ytest = train_test_split(X, y, test_size = 0.1, random_state=4)
rf = RandomForestRegressor(n_estimators=500,
                           criterion='mse',
                           random_state=4,
                           n_jobs=-1)
rf.fit(Xtrain, ytrain)
ytrain_pred = rf.predict(Xtrain)
ytest_pred = rf.predict(Xtest)
rmse_rf= (mean_squared_error(ytest,ytest_pred))**(1/2)

print('RMSE test: %.3f' % rmse_rf)
print('R^2 test: %.3f' % (r2_score(ytest, ytest_pred)))
```

```
RMSE test: 37.444
R^2 test: 0.627
```

c) Ridge Model- A ridge model was created with alpha=0.5

RidgeModel.score(Xtest,ytest) yielded a score of 0.5627088181279468

The reason behind the the poor score was attributed to the facts were the uneven distribution of p layer improvement, in that players with little improvement/decline were more common than play ers with big improvement/decline

5. Conclusions

I analysed the data for the cities Boston and Chicago and interpreted that it is more affordable to relocate to Chicago out of the two options, in terms of affordability, cost index and property index.

The crime data helped in identifying safe community areas hence the safe neighborhoods. Community Areas Edison Park, Burnside, Forest Glen, Mount Greenwood, Montclare are the top 5 safe areas The Four square API were helpful in further recommending neighborhoods out of the list in regard to common venues nearby. The selected cluster had a convenience store, gas station, bus station etc nearby ideal for commuting for living comfortably

The list of schools and residential complexes within these localities were identified and suggested for consideration.

Further using the Airbnb datasets, a predictive model was constructed for predicting the price given the set of amenities. Once the person finalizes a neighborhood depending on his needs and preferences, he can predict the prices. It was observed prices of properties in districts with many eating joints and markets were higher.

6. Future Directions

- Accuracy of the models has room for improvement
- The model can be streamlined for safer neighborhoods. With missing community codes the data can be merged with community area dataframes.
- We can further analyse crime data on types of crime committed, month wise breakup. We can load crime data from previous years and see the trends.