**Rent Pricing & Venue Data Analysis of Chicago**

**IBM Data Science Professional Certificate - Capstone Project**

Alex Johnson

June 2020

1. **Introduction**
   1. **Background**

Chicago, IL is the 3rd largest city by population in the United States of America, home to 2.7 million residents. It is comprised of 51 neighborhoods and 77 community areas covering 234 square miles of land. It is home to countless attractions such a major sporting events and world-class restaurants. Average rental housing can cost anywhere from $600 to more than $2,500 per month.

* 1. **Problem**

The size and diversity of the city can be daunting for potential immigrants, whether domestic or international. It is difficult to narrow down a neighborhood within your budget and interests. This project aims to ease this process by displaying average rent costs and popular venues within each neighborhood.

* 1. **Interest**

I’m interested in this topic because sometime in the near future I hope to intern or work in Chicago and will have to go through the housing process. Others going through this process may be interested as well.

1. **Data** 
   1. **Data Sources**

Foursquare API was used to obtain the most common venues for each neighborhood of Chicago

Rent Jungles’ rent trend database was used to analyze and display average rent prices

Spatial data from the Chicago Data Portal was used to display neighborhood boundaries on the choropleth map

Google Maps was used in order to get center coordinates of each Neighborhood

* 1. **Data Cleaning**

The majority of data cleaning was done prior to preprocessing in Microsoft Excel. Average Rent statistics were appended to neighborhood central coordinates and processed into the notebook. There were a few initial problems with the data frame. The column now labeled “Average” was previously named “ Average Rent”. The added space caused future errors.

A screenshot of a cell phone

Description automatically generated

1. **Methodology**

My GitHub repository was used as a database for this project. My master data frame contained *Neighborhood, Average Rent, Latitude,* and *Longitude* information of the city of Chicago.

In order to visualize my analysis, I utilized the python folium library. The graph below displays a map of Chicago with a marker on each neighborhood, using the latitude and longitude values.

A close up of a map

Description automatically generated

The Foursquare API database was used in order to explore popular venues in each distinct neighborhood. I designed the getNearbyVenue function to limit 100 venues within a 500 meter radius from the central latitude and longitude coordinates of each neighborhood.

The graph below displays the distribution of venues by neighborhood in Chicago. We can see central neighborhoods such as Loop, Lincoln Park, and Lincoln Square each contain more than 80 venues whereas the bottom 50 percent has less than 20 venues per neighborhood. A total of 1324 venues were found within these neighborhoods.

A screenshot of a cell phone

Description automatically generated

To summarize the venue data, I compiled a data frame of the top 10 most common venues by neighborhood.

A screenshot of a cell phone

Description automatically generated

With a brief visual analysis, we observe some clear similarities between neighborhoods. For this reason, I used k-means clustering to divide the data. The neighborhoods will be divided into non-overlapping subsets with no cluster-internal structure.

In order to determine the correct value for k, I ran the algorithm with a k value of 2 and used the elbow method to display the optimal value.

A screenshot of a cell phone

Description automatically generated

Using the elbow method, I could rule that 3 is the correct k-value and I re-ran the clustering algorithm.

Once we have identified our clusters, using pyplot, we can analyze the distribution of most common venue types by cluster.

A screenshot of a cell phone

Description automatically generated

Through the cluster venue type distribution we are able to label each cluster:

Cluster 0 – Mexican/Fast Food Restaurants

Cluster 1 – Multiple Social Venues

Cluster 2 – Parks

We can also use a histogram to display the average rent prices throughout the city. As we can see the vast majority of properties are between $1000 - $1750 per month.

A screenshot of a cell phone

Description automatically generated

To provide a thorough summary of each neighborhood on the map I used the join function to display the count of the top three venue types for each neighborhood.

A screenshot of a cell phone

Description automatically generated

I then merged those results with all related cluster and venue information to create one cumulative data frame.

A screenshot of a cell phone

Description automatically generated

A close up of a map

Description automatically generatedUsing the cumulative data frame, I recreated the folium neighborhood graph using results from the k-means clustering.

1. **Results**

A close up of a map

Description automatically generatedFinally, I used the Geojson file from the Chicago Data Portal to display the boundaries of each neighborhood. I created a choropleth map in order to overlay the Average Rent Prices by neighborhood. When a neighborhood is selected the Neighborhood Name, Cluster Name, Rent Level, and Top 3 Venues are displayed.

1. **Discussion**

Chicago, as previously noted, is an extremely large, incredibly diverse city. Yet, the range in average rent prices still shocked me, with a difference of almost $2000 per month in some neighborhoods. Also, the sheer number of venues, 1324 was quite astonishing.

I was a bit disheartened at the effectiveness of my k-means analysis. The optimal value of k at 3 left a single neighborhood in the final cluster. But, the distribution of venue types came out nicely.

In order get obtain central neighborhood coordinated I searched nearby locations for each neighborhood. This was a long, annoying process that I’m sure could be automated by a more experienced programmer.

1. **Conclusion**

As with anything, there is much deeper analysis to be done on this topic. For myself, the primary goal of this project wasn’t to see what type of restaurant were near Grant Park but to strengthen and develop the data science tools I’ve learned throughout this course. But if it helps someone find their new apartment, fantastic.

Alex Johnson