Coursera Capstone Project

Chicagos best neighborhoods for new Chinese restaurants

- Final Report -

Alexander Kowsik

June 2020

1 Purpose

This document is the final report for my peer-reviewed capstone project for the IBM Data Science Professional Certificate Program.
For any additional information about the implementation and the complete code can be found at the associated Jupyter Notebook in the GitHub repository of this project.

2 Introduction

2.1 Business Problem

Opening a new restaurant in a big city can be a very challenging endeavour. Many factors have to be considered beforehand, the location of the restaurant being the most important one. Finding a quality spot in the city significantly increases the chances for survival of the new locality as well as the number of customers, the growth and therefore also the revenue.
This project aims to give a first guideline for this decision process by looking at the example of opening a new Chinese restaurant in Chicago. The mission is to find the best neighborhoods for such a reopening by looking at statistics about crime rates, median household incomes, rent prices as well as locations of clusters of restaurants. By comparing all of the above factors a conclusion is made for the safest, cheapest, but also most lucrative neighborhoods.

2.2 Target audience

This report is useful primarily for people who want to open a new (chinese) restaurant in Chicago as it provides an analysis of datasets around the city to locate the best areas for a reopening. However, it can be of use to all people who want to open a restaurant in any city in general, since the approach that was taken in this project can be applied to other cities as well.
Because this project explores data about crimes, incomes and rent prices, it reveals a lot of insights about the the city of Chicago itself, so in a broader sense it might be of interest to people who want to find out more a Chicago.

3 Data

The following gives an overview and a description of all the datasets that were used in this project.

3.1 Chicago community areas

All of the data that was used is structured by the community areas of Chicago which have a unique number between 1 and 77 as well as a unique name. However, not every dataset contained both, so the data for the community areas had to be imported first into a pandas dataframe called chicago_data. This dataset was also used to plot the frontiers of the neighborhoods in all the maps that were created along the way. Also as a remark, throughout this report the terms community area and neighborhood are used interchangeably.
The data was retrieved from the data portal from the official homepage of Chicago, data.cityofchicago.org [2] using BeautifulSoup4. The following is an overview of the fields in the dataset:

Most important for us are the fields AREA_NUMBE as well as COMMUNITY, which are the community area code and the community name for each community area in the city.

3.2 Crime rate dataset

To find the safest neighborhoods data about crime rates is a good indicator that ca be used. Presumably not all the neighborhoods are equally affected by crimes, so data about crime rates is useful. A simple descending sort of the neighborhoods by their crime rates could reveal what neighborhoods are safe and which are not.
The data was manually retrieved from the data portal of Chicago as well, data.cityofchicago.org[3], and was imported as a pandas dataframe. This dataset contains all of the crimes from 2001 to present, but since it is a very large dataset and not all of the data is needed, only the crimes from the start of 2019 to May 2020 were used, which should be sufficient to get an overview of the crime numbers for each neighborhood.
The following is an overview of the fields in the dataset:

Column Name	Description	Type
ID	Unique identifier for the record.	Number #
Case Number	The Chicago Police Department RD Numbe...	Plain Text T
Date	Date when the incident occurred. this is so...	Date & Time
Block	The partially redacted address where the i...	Plain Text T
IUCR	The Illinois Unifrom Crime Reporting code. ...	Plain Text T
Primary Type	The primary description of the IUCR code.	Plain Text T
Description	The secondary description of the IUCR cod...	Plain Text T
Location Description	Description of the location where the incid...	Plain Text T
Arrest	Indicates whether an arrest was made.	Checkbox ✔
Domestic	Indicates whether the incident was domest...	Checkbox ✔
Beat	Indicates the beat where the incident occu...	Plain Text T
District	Indicates the police district where the incid...	Plain Text T
Ward	The ward (City Council district) where the i...	Number #
Community Area	Indicates the community area where the in...	Plain Text T
FBI Code	Indicates the crime classification as outline...	Plain Text T
X Coordinate	The x coordinate of the location where the ...	Number #
Y Coordinate	The y coordinate of the location where the ...	Number #
Year	Year the incident occurred.	Number #
Updated On	Date and time the record was last updated.	Date & Time
Latitude	The latitude of the location where the incid...	Number #
Longitude	The longitude of the location where the inc...	Number #
Location	The location where the incident occurred i...	Location

3.3 Median household income dataset

In general, wealthier people are more likely to eat out, so it would be a good idea to open the Chinese restaurant in a place where a larger number of customers is expected. Therefore data about the median household income of each neighborhood is useful. By identifying the wealthier areas the neighborhoods can be narrowed down to the more lucrative ones.
The used dataset was retrieved from homesnacks.net [5] using BeautifulSoup4. The following is an overview of the fields and the first few entries in the dataset after importing it to pandas dataframe:

	Rank	Neighborhood	Median Household Income<
0	1	Forest Glen	$112,032
1	2	Lincoln Park	$99,720
2	3	North Center	$99,384
3	4	Beverly	$99,102
4	5	Edison Park	$98,327

3.4 Rent prices dataset

For the rent prices the dataset from rentcafe.net [1] was chosen (imported using BeautifulSoup4). Although this data resembles mostly rent prices for appartments, the assumption is that it is correlated with rent prices for restaurant places. This is important, since a lot restaurants rent their place and the rent price is a major factor in the survival rate of restaurants.
The following is an overview of the fields and the first few entries in the dataset after importing it to pandas dataframe:

	Neighborhood	Average Rent
0	The Island	$562
1	Austin	$562
2	West Pullman	$612
3	Rosemoor	$612
4	Roseland	$612

3.5 Foursquare data

Data about the location of restaurants in the city can be used to determine clusters of restaurants and to identify the neighborhoods with a high density of restaurants. Opening a restaurant in one of those is probably not a good idea since the competition might be very strong. Those with very few restaurants are most likely not lucrative, so we aim at the ones in between.
Also all locations of Chinese restaurants can be retrieved and compared to the locations of all restaurants to spot the neighborhoods that have a high density of restaurants. Maybe there are

neighborhoods with a lot of restaurants, but only very few Chinese restaurants? This would mean that people eat out a lot in this areas, but not many chinese restaurants are present, so opening one there might be profitable.
To retrieve the information about all the venues and restaurant locations the API from Foursquare [4] was used. The datapoints contain information about the venue like the venue type as well as the location.
The following code displays the function that makes the API call and retrieves all the specified venues in a given area:

```
def get_nearby_venues(names, latitudes, longitudes, radius=500, query=''):

    venues_list=[]
    for name, lat, lng in zip(names, latitudes, longitudes):
        print(name)

        # create the API request URL
        url = ('https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}' \
                + '&v={}&ll={},{}&radius={}&query={}').format(
            CLIENT_ID,
            CLIENT_SECRET,
            VERSION,
            lat,
            lng,
            radius,
            query)

        # make the GET request
        results = requests.get(url).json()["response"]['groups'][0]['items']

        # return only relevant information for each nearby venue
        venues_list.append([(
            name,
            lat,
            lng,
            v['venue']['name'],
            v['venue']['location']['lat'],
            v['venue']['location']['lng'],
            v['venue']['categories'][0]['name']) for v in results])

    nearby_venues = pd.DataFrame([item for venue_list in venues_list for item in venue_list])
    nearby_venues.columns = ['Neighborhood',
                  'Neighborhood Latitude',
                  'Neighborhood Longitude',
                  'Venue',
                  'Venue Latitude',
                  'Venue Longitude',
                  'Venue Category']

    return(nearby_venues)
```

4 Methodology

4.1 Preprocessing

The first step with every dataset has been to retrieve the data from the data sources and import it into pandas dataframes, so it could be further used for analysis. After that the data was preprocessed which included:

- converting the columns to the right data type
- cleaning unnecessary information
- removing null/NaN values (e.g. crimes without location were useless in our case)
- merging of the most important data into chicago_data for plotting graphs and maps
- verifying the correctness of the resulting dataframes

One challenge was that the data for the community areas was provided in the .geojson format, which required additional processing to be used in folium to plot them on a map. For this the package geopandas came in handy.

Some of the data didn't contain all the required information, e.g. most datasets had only the area numbers of the community areas but not the names, so joins had to be performed with the chicago_data data to get all of the necessary attributes.

After all the preprocessing, the analysis of the data could begin.

4.2 Map of the community areas

To get a first feel of the location of the community areas a map with all the borders of each neighborhoods was created. This map was used in all other maps to differentiate between the individual neighborhoods. The map required the coordinates of Chicago which were provided by the geolocator package:

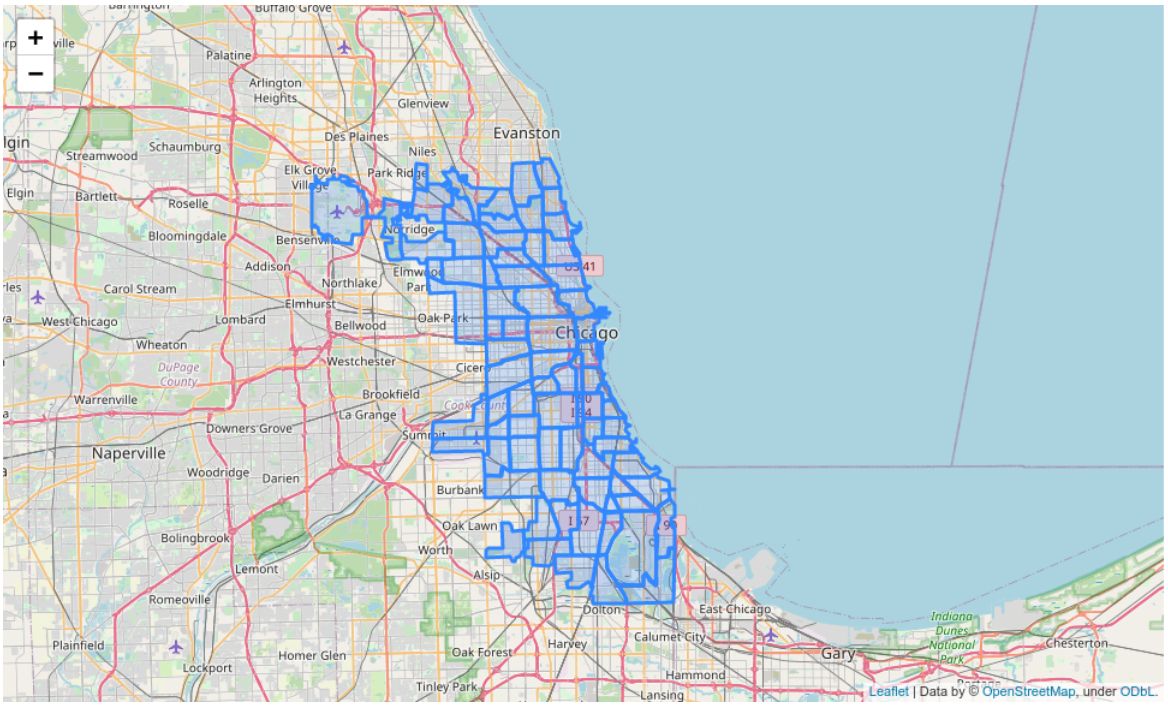
```
address = 'Chicago'

geolocator = Nominatim(user_agent="foursquare_agent")
location = geolocator.geocode(address)
latitude = location.latitude
longitude = location.longitude
print('The geograpical coordinate of Chicago are {}, {}.'.format(latitude, longitude))
```

```
The geograpical coordinate of Chicago are 41.8755616, -87.6244212.
```

The following code was used to create the map. First the polygon coordinates for each neighborhood that were located in our geojson file were imported and then the GeoJson function of folium was used to create the map.

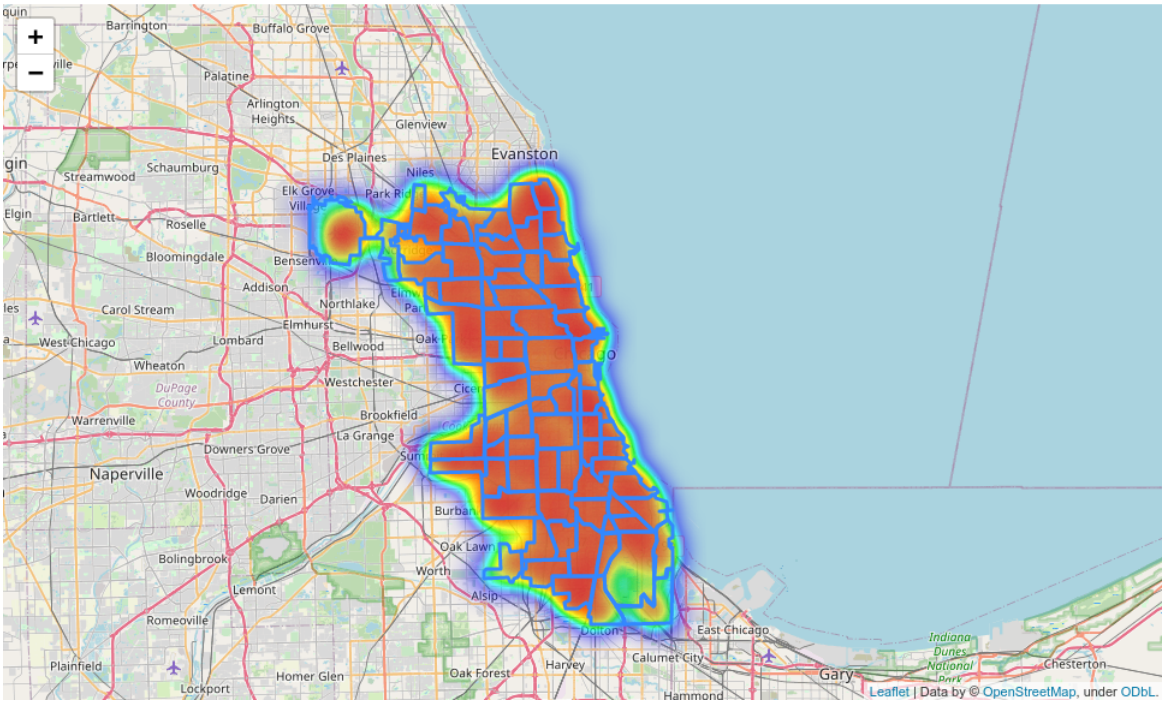
```
map_chicago = folium.Map(location=[latitude, longitude], zoom_start=10)

with open('data/Boundaries - Community Areas (current).geojson') as f:
    community_area_geodata = geojson.load(f)

folium.GeoJson(community_area_geodata).add_to(map_chicago)

map_chicago
```

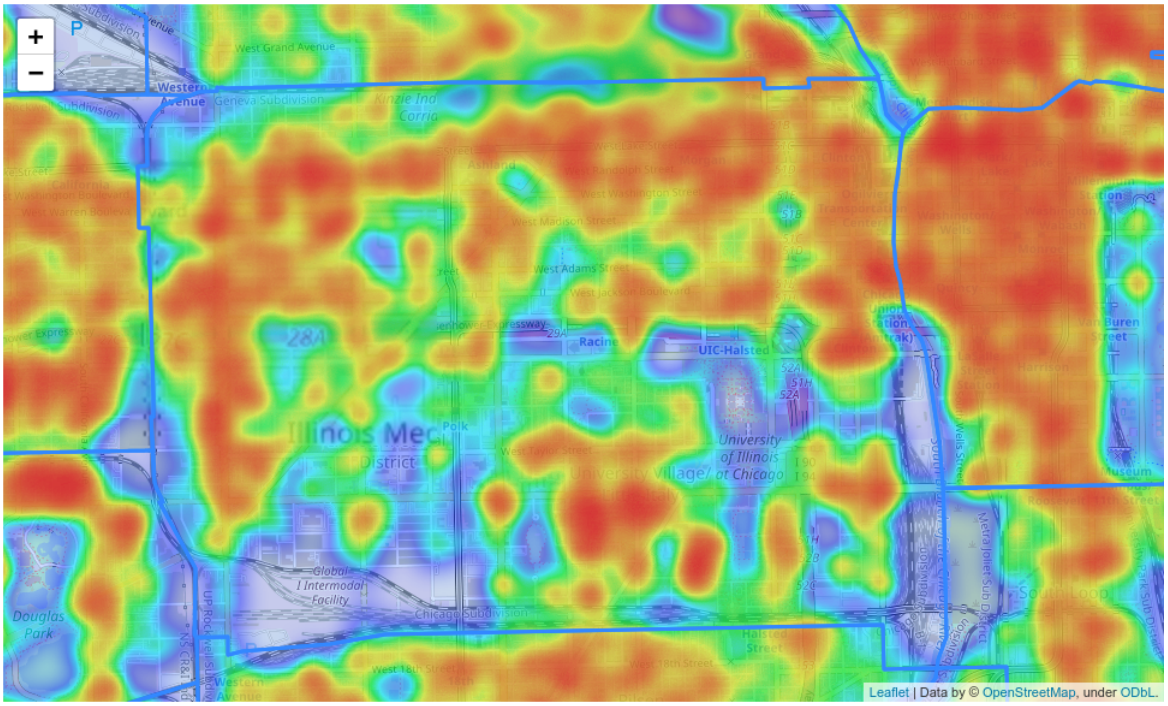
The resulting map looked like this:



4.3 Heat map of the crime rates

Before plotting any diagrams and other graphs about the data, a few additional maps were created to be able to visually spot relevant neighborhoods and get a feel for how the city is structured. A first idea was to create a heat map with the crime data, which was done with the HeatMap function of folium.plugins. The heatmap turned out to be a little bit overloaded at this zoom level and worked better with higher zoom values, both can be seen in the pictures below:


Barrington
Buffalo Grove
Palatine
Arlington Heights
Glenview
Des Plaines
Evanston
Niles
Schaumburg
Streamwood
Elk Grove Village
Park Ridge
Elgin
Bartlett
Roselle
Bloomingdale
Addison
Northlake
Carol Stream
Elmhurst
West Chicago
Lombard
Bellwood
Wheaton
Westchester
DuPage County
Warrenville
Brookfield
La Grange
Downers Grove
Naperville
Woodridge
Darien
Oak Lawn
Bolingbrook
Worth
Lemont
Alsip
Romeoville
Plainfield
Lockport
Homer Glen
Oak Forest
Harvey
Calumet City
East Chicago
Gary
Hammond
Indiana Dunes National Park
Chesterton
Leaflet | Data by © OpenStreetMap, under ODbL.


West Grand Avenue
Western Avenue
Geneva Subdivision
Kinzie Ind Corrig
West Madison Street
West Adams Street
West Jackson Boulevard
Racine
UIC-Halsted
University of Illinois at Chicago
Polk
West Taylor Street
Global I Intermodal Facility
Chicago Subdivision
Douglas Park
West 18th Street
West 16th Street
Leaflet | Data by © OpenStreetMap, under ODbL.

4.4 Color maps of all the data sets

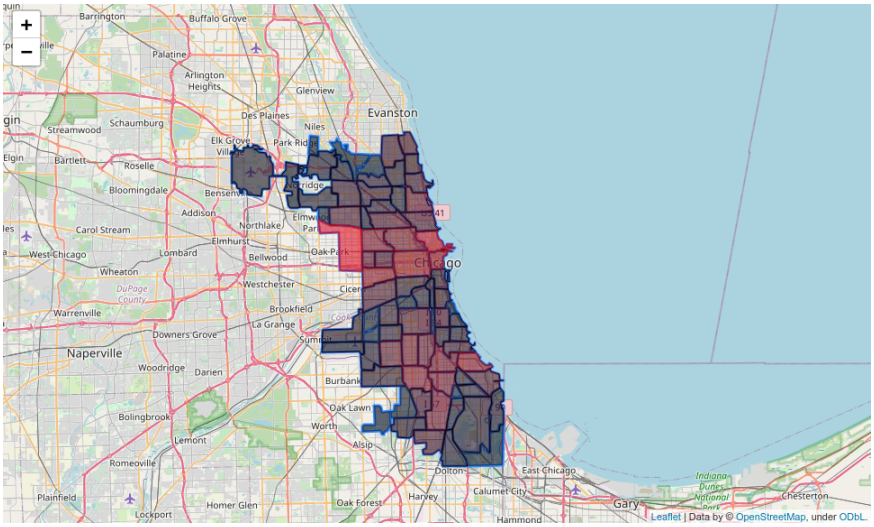
Because the heatmap was too overloaded, a different approach was made to create visuals on the map. The crime numbers, median household incomes, rent prices, restaurant numbers and the numbers of Chinese restaurants were normalized by their maximum value and colored in according to that new number. So every community area has a different shade of red, depending of its number, darker colors meaning a lower number and red colors meaning high numbers. The following is the code for generating the color and the results:

Generating the hex colors needed by folium:

```
maximum = counted_chinese_restaurants['Venue'].max()
counted_chinese_restaurants['chinese_restaurant_colors'] = counted_chinese_restaurants['Venue']\
                                                    .apply(lambda x: convert_to_hex(x/maximum)).tolist()
counted_chinese_restaurants['chinese_restaurant_colors'][0:10]
```

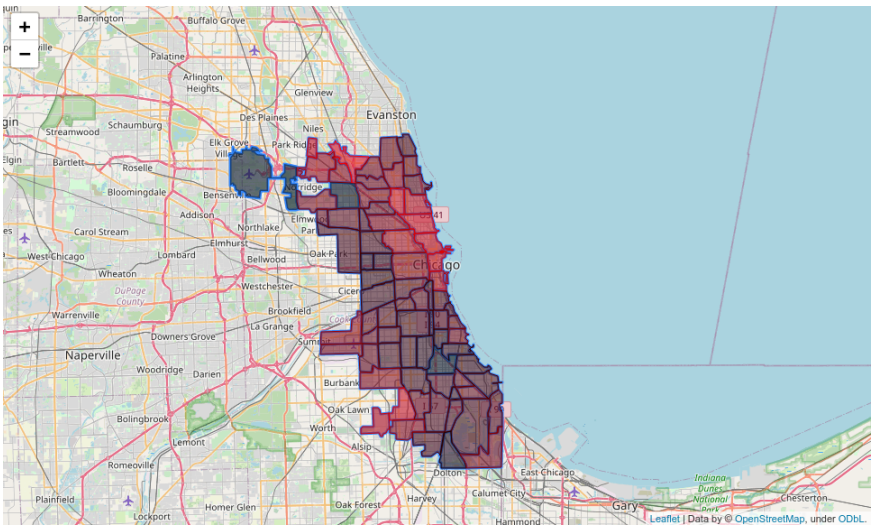
```
Neighborhood
ALBANY PARK       #3f0000
ARMOUR SQUARE     #9f0000
AVONDALE          #3f0000
BELMONT CRAGIN    #1f0000
BRIDGEPORT        #Ff0000
CLEARING          #1f0000
DOUGLAS           #1f0000
EAST SIDE         #1f0000
EDGEWATER         #1f0000
ENGLEWOOD         #1f0000
Name: chinese_restaurant_colors, dtype: object
```

Color map of the crime numbers:



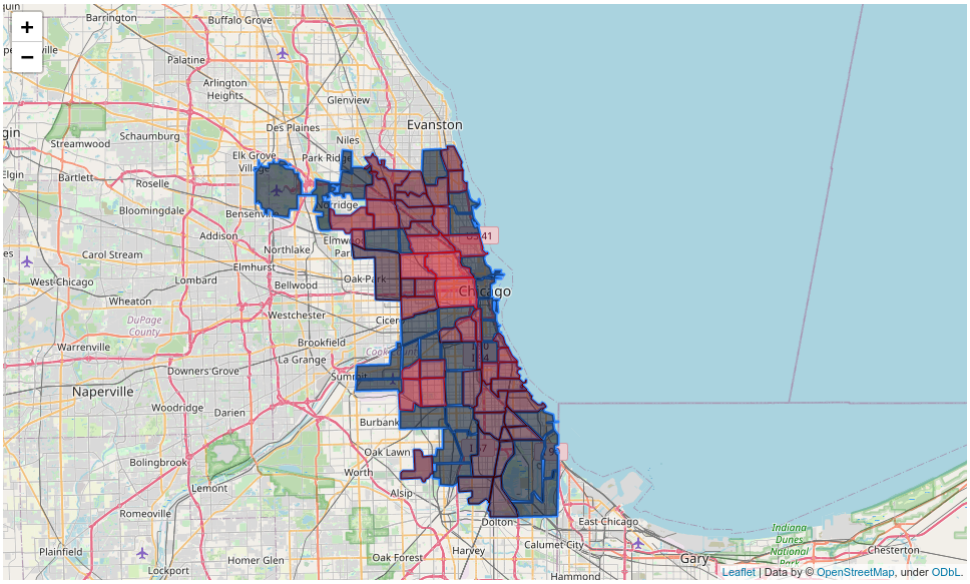
In this view it can be differentiated between the crime numbers in each community area much better. The crime numbers are especially high in the middle of the lower half of the city, as well as in its center, reaching a visible maximum in the west of the center.

Color map of the median household incomes:



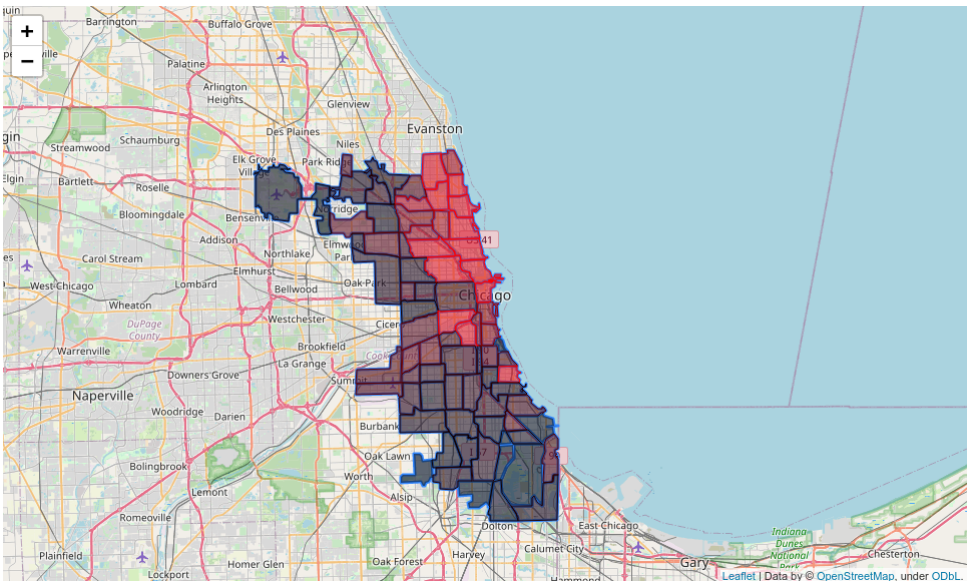
It can be observed that the wealthier neighborhoods lie in the north east and south west of the city.

Color map of the rent prices:



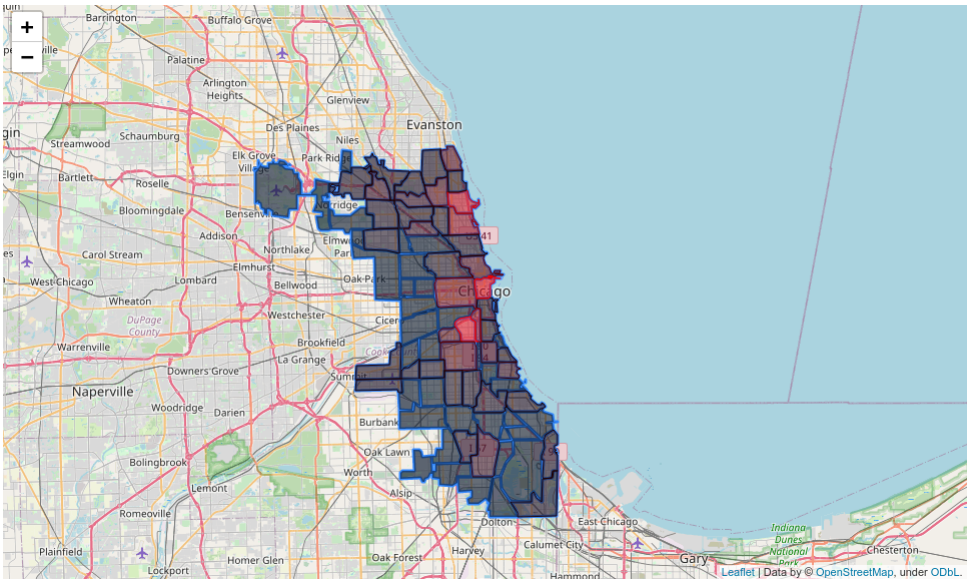
At the first sight it is clear that opening a restaurant in the middle of the city can get quite costly.

Color map of the restaurant numbers in each neighborhood:



Interesting! Most of the restaurants are located in the north east of the city.

Color map of numbers of Chinese restaurants in each neighborhood:



As a lot of the neighborhoods are pretty grey it can be concluded that most of the Chinese restaurants are concentrated in only a few neighborhoods in the center and the north west area of the city. This correlates with the overall numbers of all restaurants.

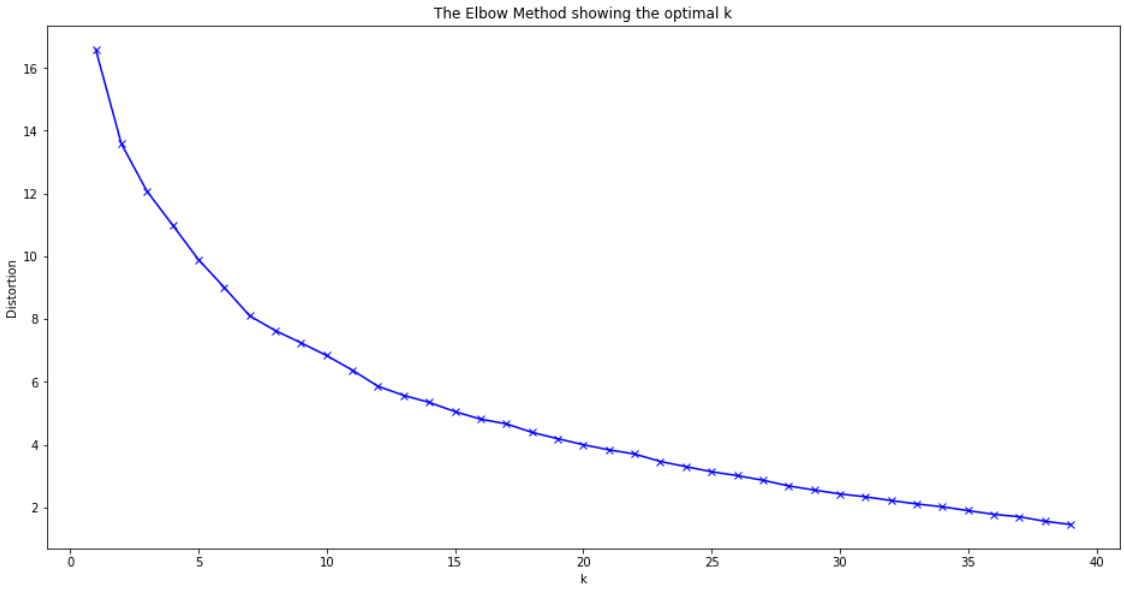
4.5 Clustering the restaurants

Before comparing the datasets among themselves, a k-means clustering algorithm was run on the restaurant datasets to see what insights might come from this.
To do this a one hot encoding of the restaurants data had to be made and the data had to be grouped together for each neighborhood. This way the mean percentage for each restaurant category and each neighborhood could be obtained and the clustering algorithm could detect clusters. Furthermore only the top ten venue types were kept to reduce the noise.

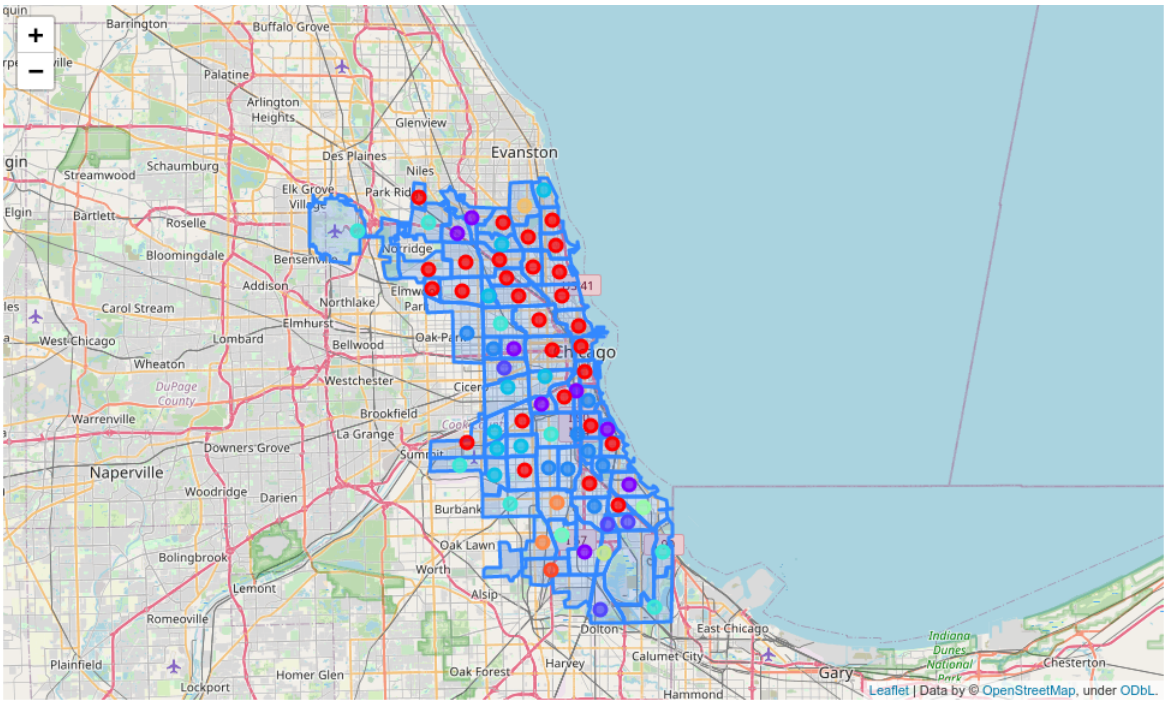
```
num_top_venues = 10

indicators = ['st', 'nd', 'rd']

# create columns according to number of top venues
columns = ['Neighborhood']
for ind in np.arange(num_top_venues):
    try:
        columns.append('{}{} Most Common Venue'.format(ind+1, indicators[ind]))
    except:
        columns.append('{}th Most Common Venue'.format(ind+1))

# create a new dataframe
restaurants_sorted = pd.DataFrame(columns=columns)
restaurants_sorted['Neighborhood'] = restaurants_grouped['Neighborhood']

for ind in np.arange(restaurants_grouped.shape[0]):
    restaurants_sorted.iloc[ind, 1:] = return_most_common_venues(
        restaurants_grouped.iloc[ind, :],
        num_top_venues)

restaurants_sorted.head()
```

	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	ALBANY PARK	Mexican Restaurant	Korean Restaurant	Sushi Restaurant	Bakery	Fast Food Restaurant	Chinese Restaurant	Latin American Restaurant	Taco Place	Asian Restaurant	BBQ Joint
1	ARCHER HEIGHTS	Mexican Restaurant	Bakery	Italian Restaurant	Seafood Restaurant	Hot Dog Joint	Pizza Place	Creperie	Cuban Restaurant	Deli / Bodega	Dim Sum Restaurant
2	ARMOUR SQUARE	Chinese Restaurant	Asian Restaurant	Bakery	Pizza Place	Mexican Restaurant	Dim Sum Restaurant	Restaurant	Sandwich Place	Indian Restaurant	Seafood Restaurant
3	ASHBURN	Pizza Place	Food	Italian Restaurant	Wings Joint	Eastern European Restaurant	Creperie	Cuban Restaurant	Deli / Bodega	Dim Sum Restaurant	Diner
4	AUBURN GRESHAM	Bakery	Fast Food Restaurant	Seafood Restaurant	Wings Joint	Eastern European Restaurant	Creperie	Cuban Restaurant	Deli / Bodega	Dim Sum Restaurant	Diner

To find the optimal number of clusters k an elbow curve was plotted. This graph has k on the x-axis and the distortion for each run of the k-means algorithm with k as parameter on the y-axis. The point where the graph starts to drastically change its slope is called the elbow point and marks the optimal k for the algorithm. In this case the optimal k was around 12. This was the elbow curve:



Finally a k-means clustering algorithm could be run for the data, the resulting clusters were plotted on a map:



It can be observed that the long red cluster in the north east of the city correlates with the high number of restaurants in this area. Other than this correlation no more useful information could be extracted from this clustering. The data from the other datasets seems to be more relevant.

4.6 Further analysis

Maps are a great tool to get a first feel for the data and to build an intuition about where to look next to gain more insights. But to get concrete results one has to look at the actual numbers. Also plotting these can further widen the understanding of the data.

To see which neighborhoods are affected by crimes the most a bar chart with descending values for crime numbers was plotted.

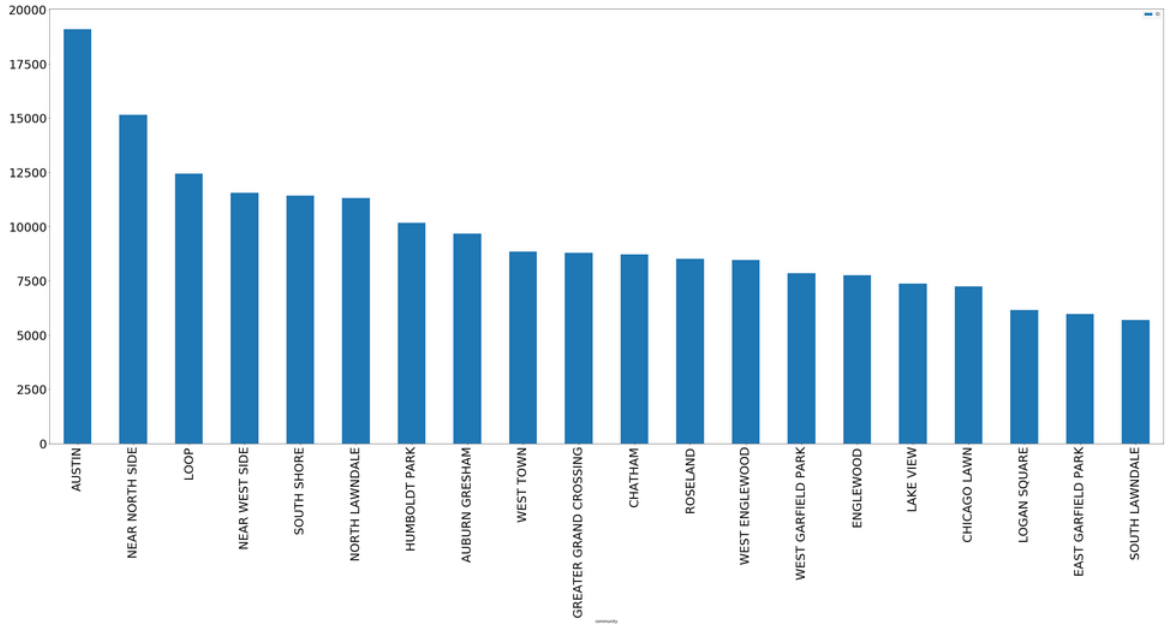
```
counted_crimes.head()
```

community	ID	crime_colors
ALBANY PARK	2869	#260000
ARCHER HEIGHTS	1048	#D0000
ARMOUR SQUARE	1318	#110000
ASHBURN	2988	#270000
AUBURN GRESHAM	9669	#810000

To see which neighborhoods are affected by crimes the most we plot a bar chart with descending values for crimes. Those numbers can be a huge factor for deciding where to open a restaurant to minimize the risk of robberies and to get in an overall safe neighborhood.

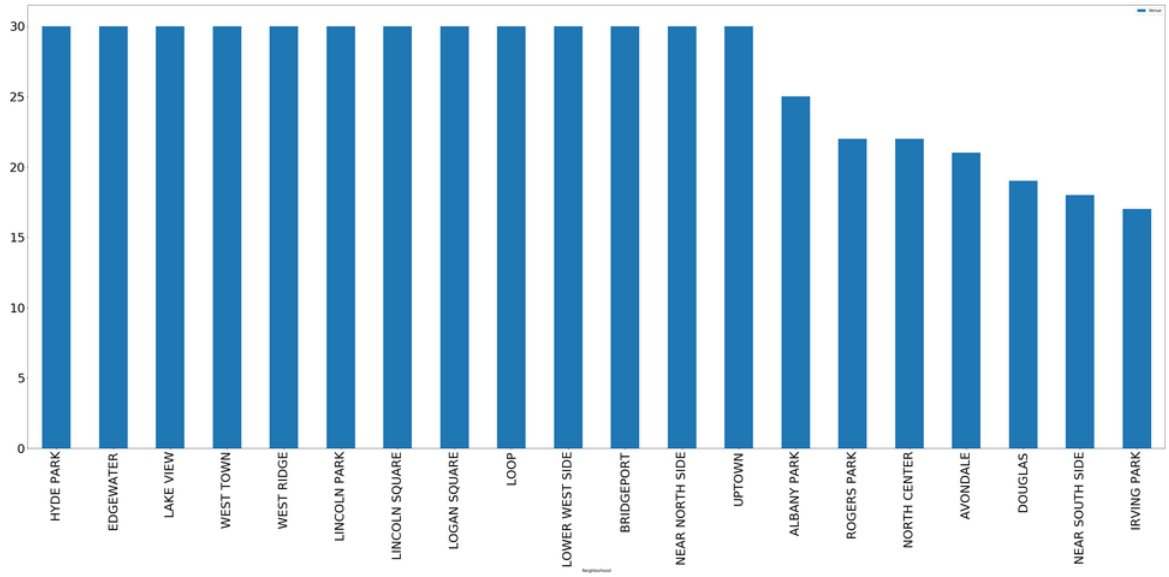
```
counted_crimes[['ID']].sort_values(by='ID', ascending=False).head(20) \
    .plot(kind='bar', figsize=(50, 20), fontsize=30)
```

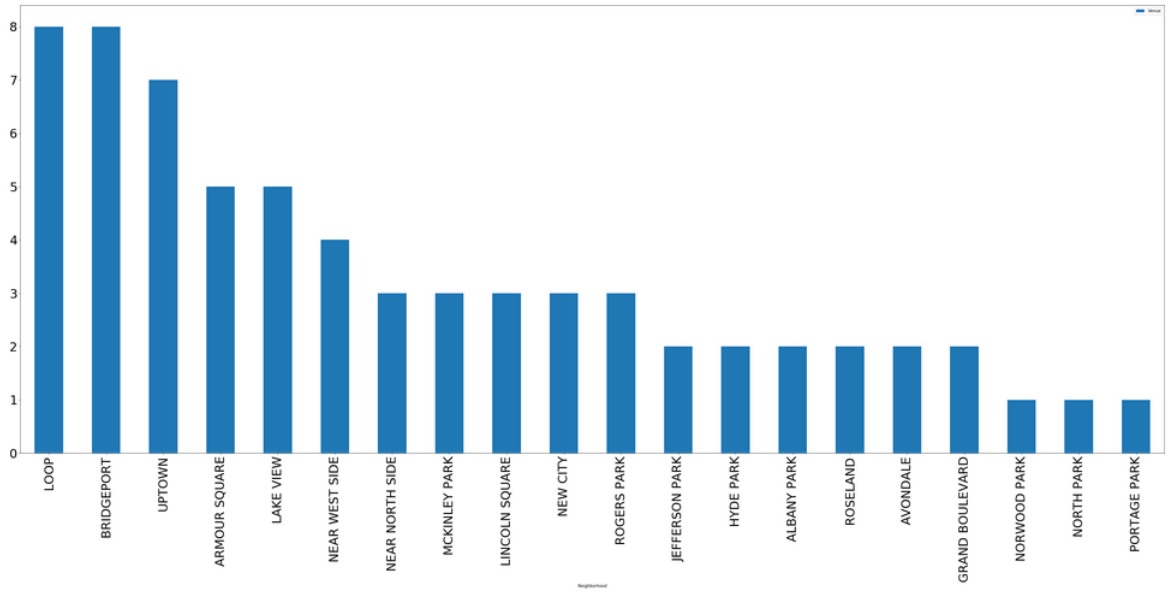
Those numbers can be a huge factor for deciding where to open a restaurant to minimize the risk of robberies and to get in an overall safe neighborhood. The following is a plot of the top 20 neighborhoods with high crime rates:



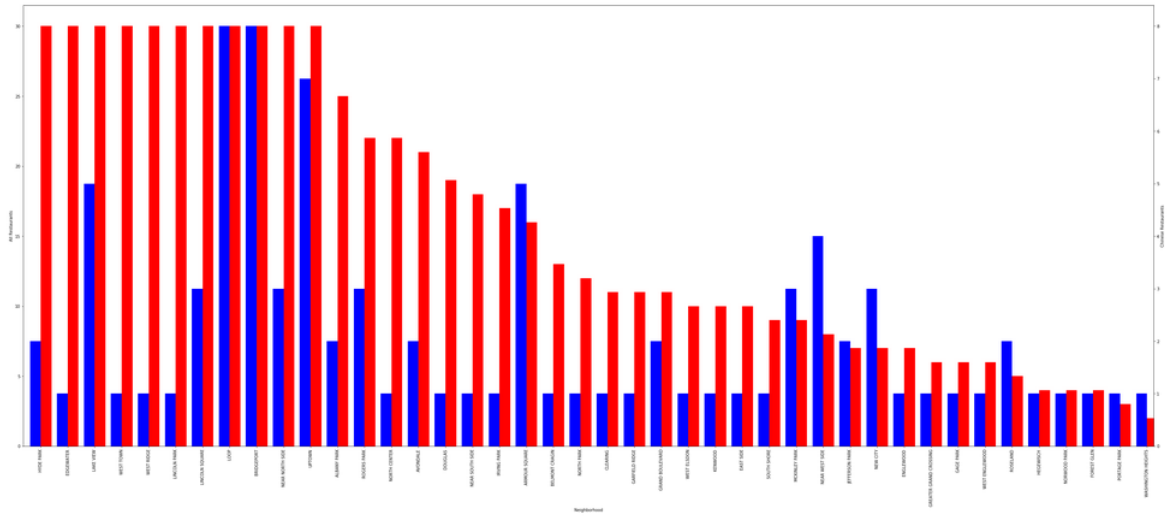
The most unsafe community areas are Austin, Near North Side, Loop, Near West Side, South Shore and North Lawndale. Picking these neighborhoods is probably not a good idea.

The same was done for the restaurant data and also the data for the chinese restaurants, resulting in the following plots:



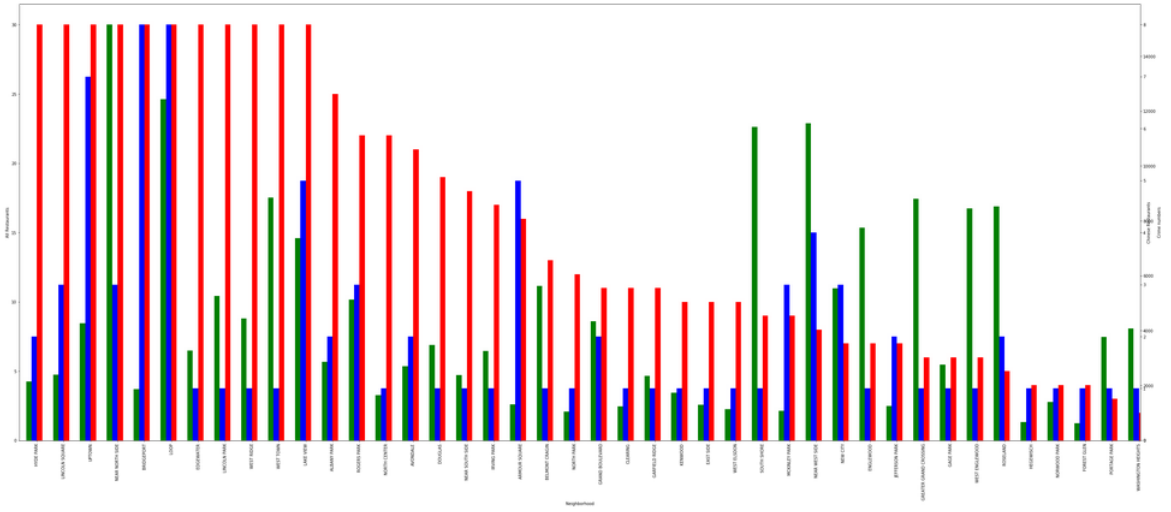


Interestingly there are some neighborhoods with a high density of restaurants but with very few Chinese restaurants which could indicate a possible gap in the market. So the next step was to plot both data in one plot (red: all restaurants, blue: Chinese restaurants):



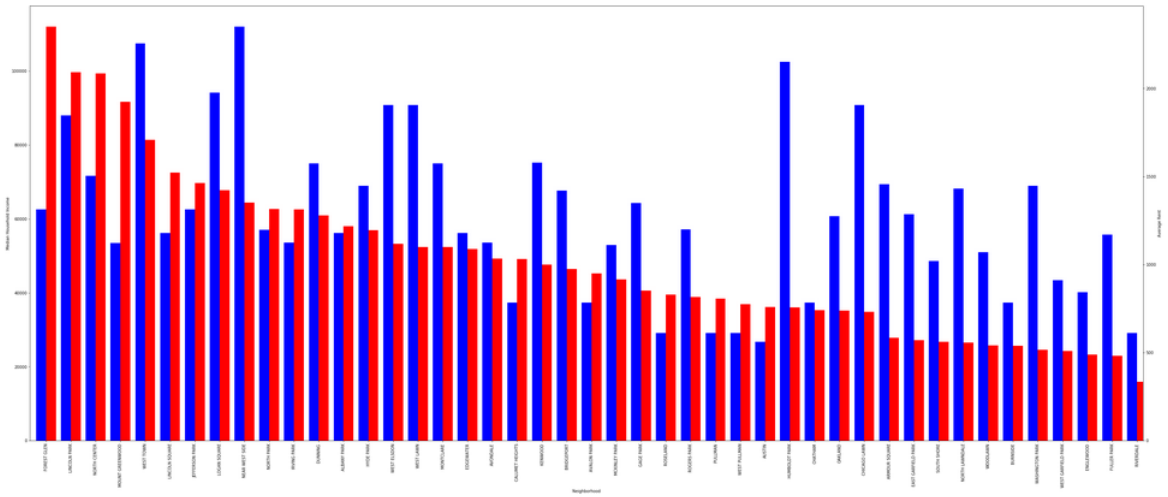
It can be observed that the neighborhoods Hydepark, Edgewater, Lake View, West Town, West Ridge, Lincoln Park and Lincoln Square don't have as much Chinese restaurants but have a high number of other restaurants, so these neighborhoods should be taken into the consideration.

The next step was to add a plot for the crime numbers for each neighborhood:



Luckily, apart from Near North Side, the areas with high crime rates are mostly in neighborhoods with very few restaurants, which are not of interest anyway.

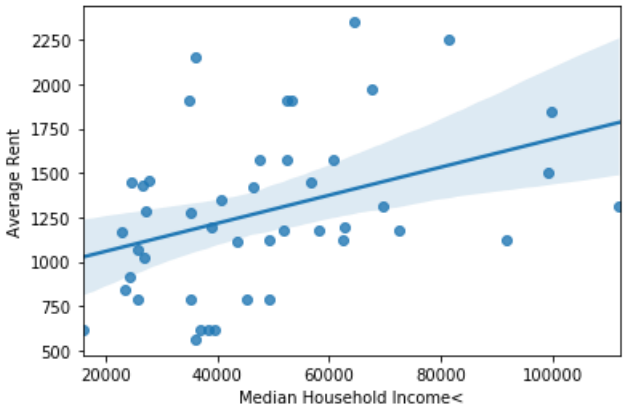
The same plots were done for the median household income and the average rent prices:



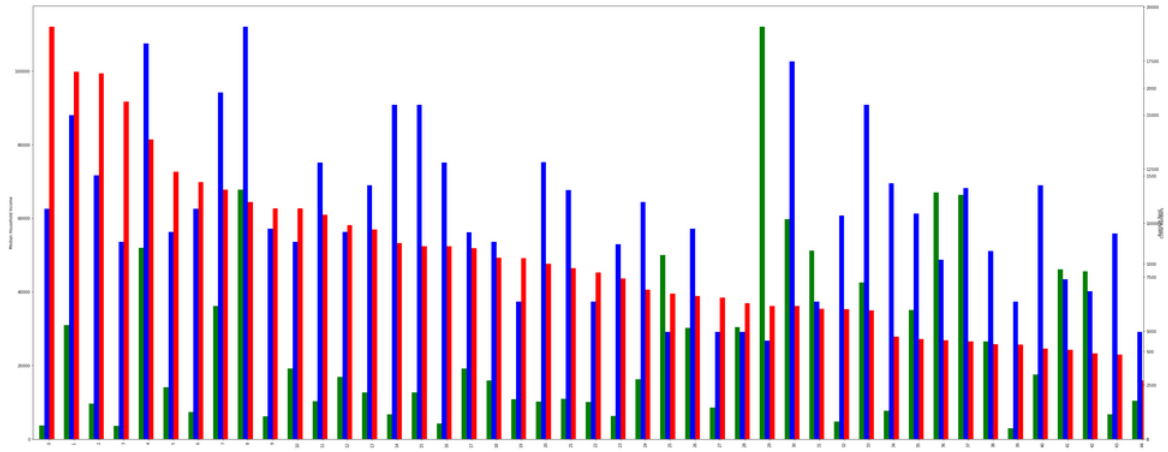
The regression plot revealed as expected a positive correlation between the household income and rent prices, so neighborhoods with wealthier people are also more expensive in regards to rent prices.

```
sns.regplot(x="Median Household Income<", y="Average Rent", data=wealth_data)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f68189084f0>
```



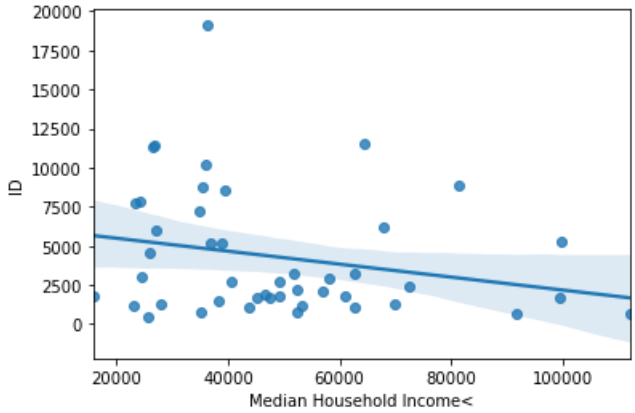
Another interesting question that was explored is whether there is a correlation between wealth and crime numbers. To see this a bar plot was created:



It seems that apart from very few exceptions the crime rates go up as the income of the citizens decreases. A correlation plot with a regression line revealed that there is indeed a negative correlation between income and crime numbers.

```
sns.regplot(x="Median Household Income<", y="ID", data=wealth_and_crimes)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f6818904f70>
```



5 Results

The results are to some degree subject to interpretation. With data about crime rates, the median household income, the rent prices and clusters of restaurants one has to to decide which factors are more relevant for himself.
What can be said objectively is that wealthier neighborhoods which are mostly located in the north east and south west of the city are correlated with lower crime rates, making them a very attractive consideration. The wealthier the neighborhood the more likely it is that people eat out and therefore more customers can be expected and the lower the crime rate, the less one has to worry about robberies. The crime rates are also located mostly in the west side of the city center.
Interestingly the rent prices only loosely correlate with the household income of the neighborhood and are high in the city center. So concentrating on the north east and south west part of the city

is the way to go if rent prices are a major concern.
Also most of the restaurants are located in the north east part of the city which can be beautifully seen in the map plots. The data for the chinese restaurants reveals that most of the neighborhood in this area have only few Chinese restaurants which raises the question if it might be lucrative to open one in this area.
So averaging all of the factors leads to the conclusion that the community areas in the north east part of the city are the ones to go for when opening a new Chinese restaurant. Some examples are all neighborhoods of the group 'North Side' and 'Far North Side' as well as 'Central'.

6 Discussion

Of course this approach can only give a first guideline to choosing a location for the reopening of a new restaurant. Many other factors have to be taken into consideration that could be important, like the possibilities for logistic, the connection to stations of public transportation, the vicinity to other restaurants, the ethnicity of the citizens in that area and many many other.
However, the big factors considerations were targeted in this project and this led to a pretty good result. But still for the final decision one has to consult other opinions and acquire more data. Maybe it is a better idea to open in the south west part of the city which was attractive as well, maybe because the opportunities there are better or the competition is lower. These types of questions are not in the scope of this project and are let to the point of view of the reader.

7 Conclusion

This concludes this report and the final capstone project of the program. It was a great journey and this course has enabled me to not only learn all of the fundamentals of Data Science and Machine Learning but also to apply my knowledge to real world problems and to work with real world data. This experience is something I am very grateful for and hope that it leads to further developments and opportunities in this area.

Alexander Kowsik

Feel free to connect with me on LinkedIn under www.linkedin.com/in/alexander-kowsik

References

[1] Average rent in chicago, il by neighborhood, https://www.rentcafe.com/average-rent-market-trends/us/il/chicago, 2020.

[2] Boundaries - community areas (current), https://data.cityofchicago.org/facilities-geographic-boundaries/boundaries-community-areas-current-/cauq-8yn6, 2020.

[3] Crimes from 2001 to present, https://data.cityofchicago.org/public-safety/crimes-2001-to-present/ijzp-q8t2, 2020.

[4] Foursquare, https://www.foursquare.com, 2020.

[5] Richest neighborhoods in chicago, il for 2020, https://www.homesnacks.net/richest-neighborhoods-in-chicago-128972, 2020.