Capstone Project - The Battle of Neighborhoods

Evangelos Patsourakos

June 7, 2020

Abstract

Many real estate firms have long made decisions based on a combination of intuition and traditional, retrospective data. Today, a host of new variables make it possible to paint more vivid pictures of a location's future risks and opportunities. In this report, we will find the correlation between the neighbourhoods of Toronto, Canada and the nearby venues and how the places influence the price of the unit.

1 Introduction and Overview

1.1 Target audience

The audience that will benefit from this analysis will be:

1. People like families that were there looking to move to Toronto.
2. Real estate companies which are looking for relations in the market.
3. Landlords that the looking to sell their properties and they are looking for an estimation.
4. Individual buyers that they are searching for a property to buy.

2 Data usage

2.1 Data gathering

To collect data on the values of the real estate units, we must use an API. One such is the Zoocasa API which is available for free. Then we will gather the data for the locations of the shops. So we will import the packages and keys for our connection to the API.

The Zoocasa API was selected because it is efficient and its availability. And then, we loop through the links to create a data frame: Figure 1, Figure 2,

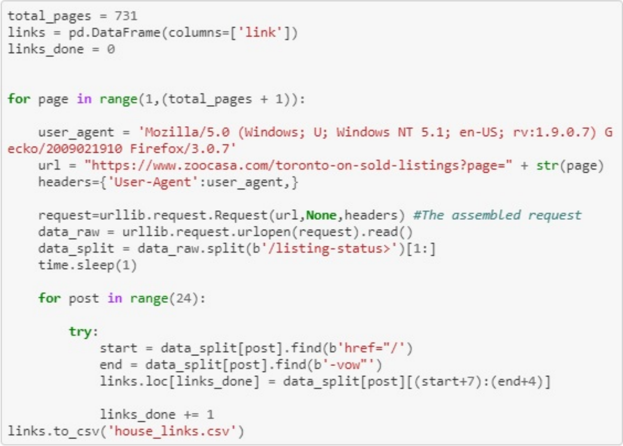
2.2 Data prepossessing

To compare the number of the venue with the real estate units we need:

- Real estate price - The square meters of the apartments - The coordinates to examine the places and the flats in a specific radius.

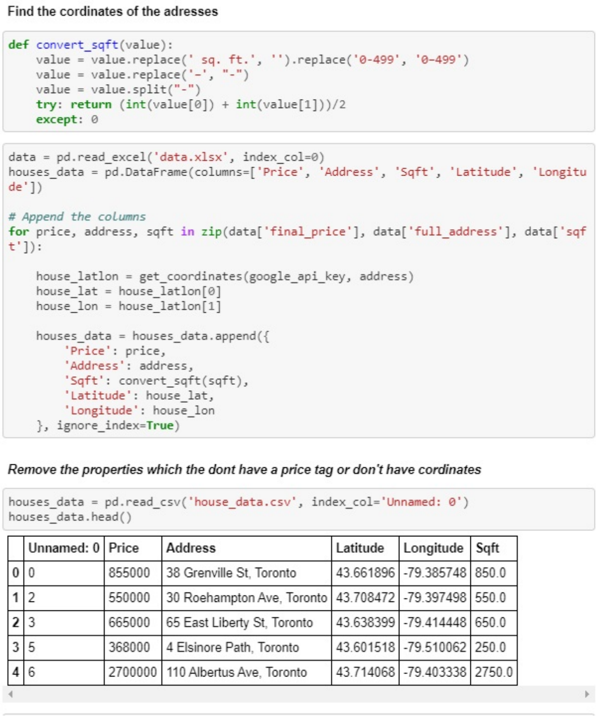
The process and cleaning of the data, we need to make the columns and the data in the appropriate format.

In the end, we get the venues for each property in a radius of 500, and we print the result on a map. [Figure 3, Figure 4,]



```
total_pages = 731
links = pd.DataFrame(columns=['link'])
links_done = 0


for page in range(1,(total_pages + 1)):

    user_agent = 'Mozilla/5.0 (Windows; U; Windows NT 5.1; en-US; rv:1.9.0.7) G
ecko/2009021910 Firefox/3.0.7'
    url = "https://www.zoocasa.com/toronto-on-sold-listings?page=" + str(page)
    headers={'User-Agent':user_agent,}

    request=urllib.request.Request(url,None,headers) #The assembled request
    data_raw = urllib.request.urlopen(request).read()
    data_split = data_raw.split(b'/listing-status>')[1:]
    time.sleep(1)

    for post in range(24):

        try:
            start = data_split[post].find(b'href="/')
            end = data_split[post].find(b'-vow"')
            links.loc[links_done] = data_split[post][(start+7):(end+4)]

            links_done += 1
links.to_csv('house_links.csv')
```

Figure 1: Code of gathering the links.



Find the cordinates of the adresses

```
def convert_sqft(value):
    value = value.replace(' sq. ft.', '').replace('0-499', '0-499')
    value = value.replace('-', "-")
    value = value.split("-")
    try: return (int(value[0]) + int(value[1]))/2
    except: 0
```

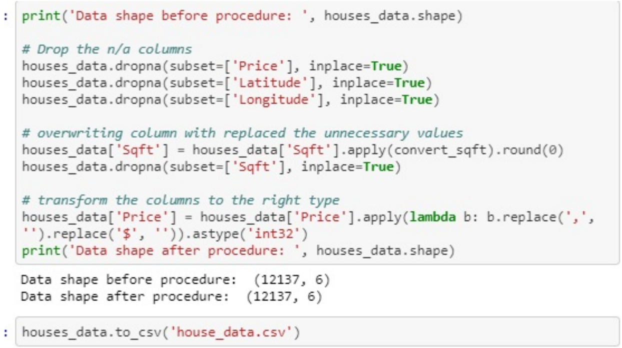
```
data = pd.read_excel('data.xlsx', index_col=0)
houses_data = pd.DataFrame(columns=['Price', 'Address', 'Sqft', 'Latitude', 'Longitu
de'])

# Append the columns
for price, address, sqft in zip(data['final_price'], data['full_address'], data['sqf
t']):

    house_latlon = get_coordinates(google_api_key, address)
    house_lat = house_latlon[0]
    house_lon = house_latlon[1]

    houses_data = houses_data.append({
        'Price': price,
        'Address': address,
        'Sqft': convert_sqft(sqft),
        'Latitude': house_lat,
        'Longitude': house_lon
    }, ignore_index=True)
```

Remove the properties which the dont have a price tag or don't have cordinates

```
houses_data = pd.read_csv('house_data.csv', index_col='Unnamed: 0')
houses_data.head()
```

	Unnamed: 0	Price	Address	Latitude	Longitude	Sqft
0	0	855000	38 Grenville St, Toronto	43.661896	-79.385748	850.0
1	2	550000	30 Roehampton Ave, Toronto	43.708472	-79.397498	550.0
2	3	665000	65 East Liberty St, Toronto	43.638399	-79.414448	650.0
3	5	368000	4 Elsinore Path, Toronto	43.601518	-79.510062	250.0
4	6	2700000	110 Albertus Ave, Toronto	43.714068	-79.403338	2750.0

Figure 2: Get the coordinates for a specific property.



```
: print('Data shape before procedure: ', houses_data.shape)

  # Drop the n/a columns
  houses_data.dropna(subset=['Price'], inplace=True)
  houses_data.dropna(subset=['Latitude'], inplace=True)
  houses_data.dropna(subset=['Longitude'], inplace=True)

  # overwriting column with replaced the unnecessary values
  houses_data['Sqft'] = houses_data['Sqft'].apply(convert_sqft).round(0)
  houses_data.dropna(subset=['Sqft'], inplace=True)

  # transform the columns to the right type
  houses_data['Price'] = houses_data['Price'].apply(lambda b: b.replace(',',
  '').replace('$', '')).astype('int32')
  print('Data shape after procedure: ', houses_data.shape)

  Data shape before procedure:  (12137, 6)
  Data shape after procedure:  (12137, 6)

: houses_data.to_csv('house_data.csv')
```

Figure 3: Clear the property data.

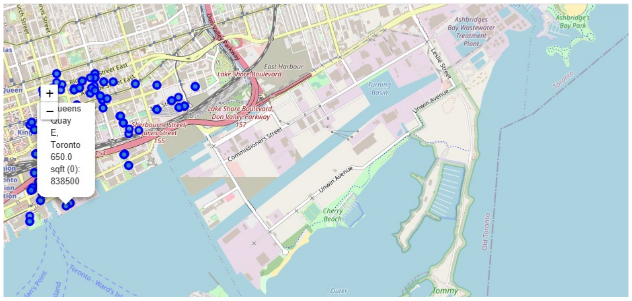


Figure 4: Toronto map.

3 Methodology

The assumption is that real estate price is correlated on the surrounding venue. Therefore, if we make a correlation analysis and try to make a model, we will be able to determinate whether it is true or false.

First, I do some visualization techniques such as Tree Repressor and heatmap. Afterwards, I go for a more mathematical approach.

In the analysis, I try to make techniques first and then construct my model.

4 Analysis

To determine which features are most important for our analysis, I build a matrix. Therefore, to estimate the importance of features, here I use Tree Repressor.

This estimation shows us that the square meter is more important than the number of shops. Moreover, the location of the flat has a significant role in the determination of the price.

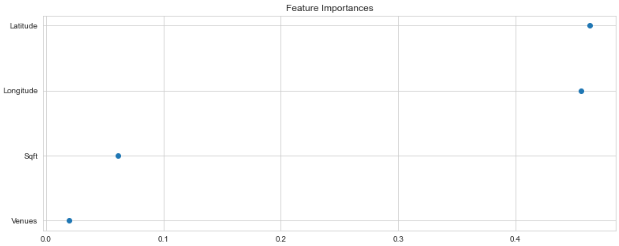


Figure 5: Tree Repressor.

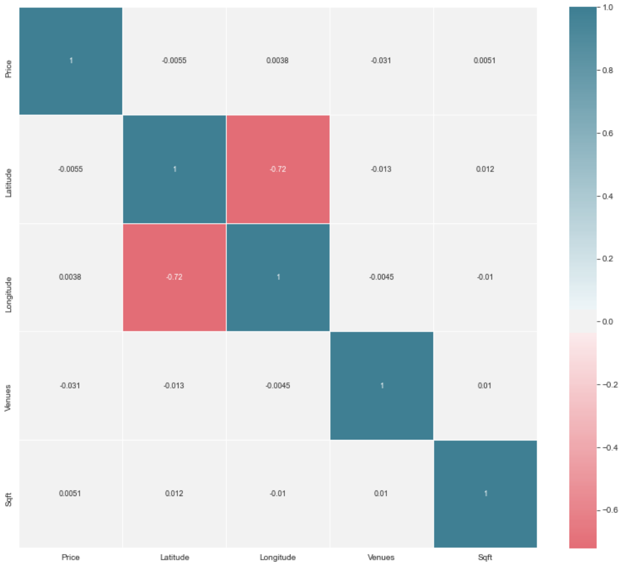
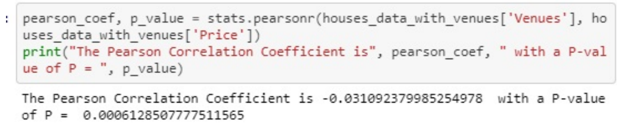


Figure 6: Heat-map to inspect the correlation.



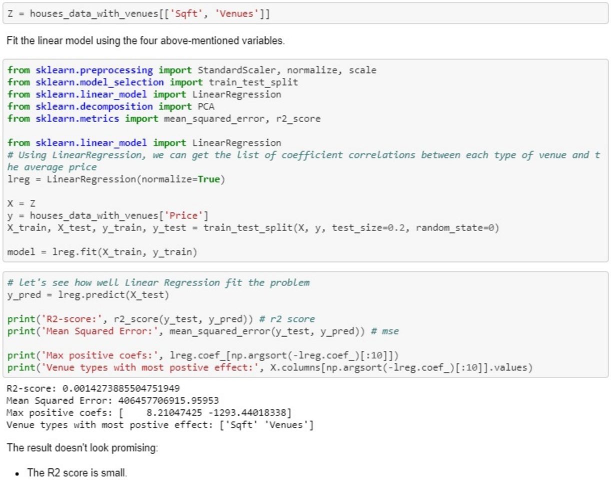
```
: pearson_coef, p_value = stats.pearsonr(houses_data_with_venues['Venues'], ho
  uses_data_with_venues['Price'])
  print("The Pearson Correlation Coefficient is", pearson_coef, " with a P-val
  ue of P = ", p_value)
```

```
The Pearson Correlation Coefficient is -0.031092379985254978  with a P-value
of P =  0.0006128507777511565
```

Figure 7: P-Value to inspect the correlation.

Regarding the correlation techniques, the p-value is ¡ 0.001, which illustrates strong evidence that the correlation is significant.

Finally, we created a linear regression model based on square meters and the number of venues. However, the evaluation gave us a small R2 score which is not that promising.



```
Z = houses_data_with_venues[['Sqft', 'Venues']]
```

Fit the linear model using the four above-mentioned variables.

```
from sklearn.preprocessing import StandardScaler, normalize, scale
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.decomposition import PCA
from sklearn.metrics import mean_squared_error, r2_score

from sklearn.linear_model import LinearRegression
# Using LinearRegression, we can get the list of coefficient correlations between each type of venue and t
he average price
lreg = LinearRegression(normalize=True)

X = Z
y = houses_data_with_venues['Price']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=0)

model = lreg.fit(X_train, y_train)
```

```
# let's see how well Linear Regression fit the problem
y_pred = lreg.predict(X_test)

print('R2-score:', r2_score(y_test, y_pred)) # r2 score
print('Mean Squared Error:', mean_squared_error(y_test, y_pred)) # mse

print('Max positive coefs:', lreg.coef_[np.argsort(-lreg.coef_)[:10]])
print('Venue types with most postive effect:', X.columns[np.argsort(-lreg.coef_)[:10]].values)
```

```
R2-score: 0.0014273885504751949
Mean Squared Error: 406457706915.95953
Max positive coefs: [    8.21047425 -1293.44018338]
Venue types with most postive effect: ['Sqft' 'Venues']
```

The result doesn't look promising:

- The R2 score is small.

(1)

5 Results

The analysis showed us that the most crucial role in the price of a real estate is the location. In the sequel, the two features with the most considerable influence were the square meters and in the end the venues. Even if the sites were last, the analysis shows that there is a strong correlation between the price and the number of stores. And then, our target group should consider this factor when they are up to buy a real estate unit.

6 Discussion

The purpose of this analysis was to find out if there was a correlation between the price of housing and the number of venues around it, to help the stakeholders the search for the best location for a new property. The way I tried to support this fact was price correlation tools and ratios between the features.