IBM – Coursera

Data Science Specialization

Capstone project - Final report

Effect of Neighborhood Venue Composition to Housing Price

11/22/2019

Table of content:

I. Introduction:

The main goal will be exploring the neighborhoods of cities of south Florida in order to extract the correlation between the real estate value and its surrounding venues.

Housing price varies dramatically depending on the house’s neighborhood. Knowing the influencers of housing price is critical to buyers and investors. Many studies already discussed impact from school zones, crime rates et.al. This study reveals the impact from neighborhood venue composition

The venue composition is the percentage of venues belongs to a category (e.g. 3% venues in Boca Raton, Fl is “American Restaurant” vs 6% in Boynton Beach, Fl)

II. Data description:

South Florida city neighborhoods were chosen as the observation target due to the following reasons:

- The availability of real estate prices. Though very limited.
- The diversity of prices between neighborhoods.
- The availability of geo data which can be used to visualize the dataset onto a map.

The type of real estate to be considered is 4-bedroom family house, which is common for most normal families.

The dataset will be composed from the following main sources:

- House Price
  - Median price by neighborhoods in US. (https://www.zillow.com/research/data/)
- Venue Data
  - FourSqaure API (https://developer.foursquare.com/)
- Geolocation Data
  - Geopy (https://geopy.readthedocs.io/en/stable/)

The process of collecting and clean data:

- Download the median price dataset from Zillow.com
- Find the geographic data of the neighborhoods.
- For each neighborhood, pass the obtained coordinates to FourSquare API. The “explore” endpoint will return a list of surrounding venues in a pre-defined radius.
- Count the occurrence of each venue type in a neighborhood. Then apply one hot encoding to turn each venue type into a column with their occurrence as the value.
- Standardize the average price by removing the mean and scaling to unit variance.

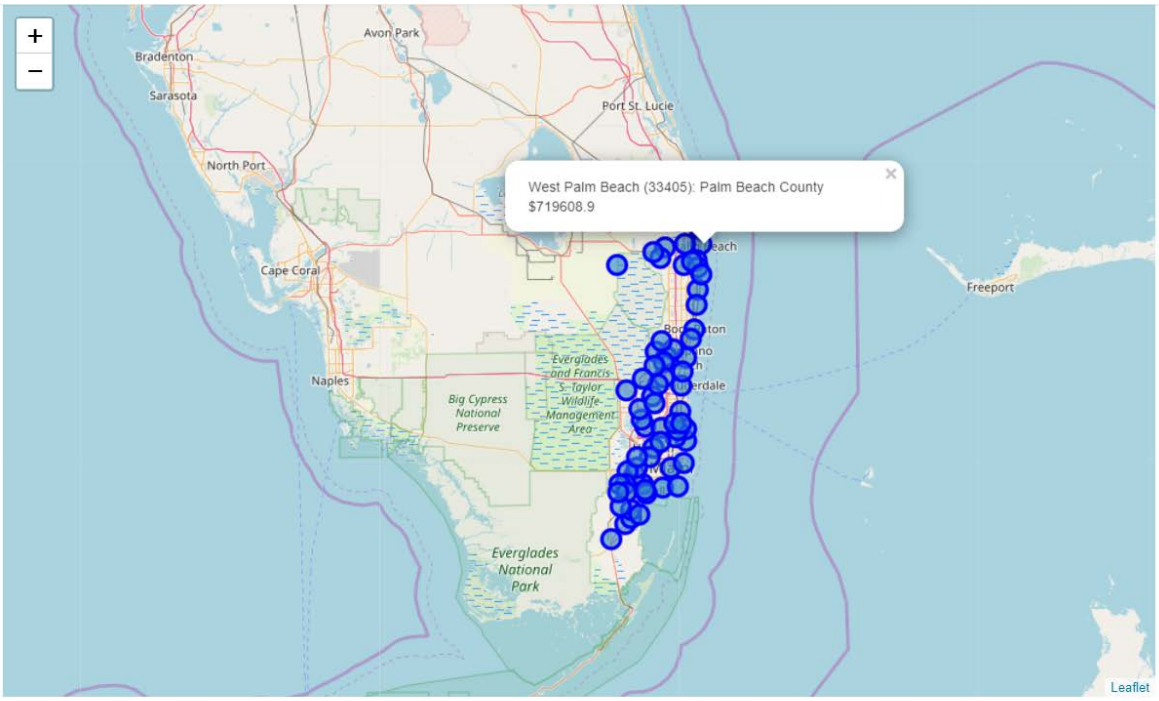
The result dataset is a 2 dimensions data frame

- Each row represents a neighborhood.

- Each column, except the last one, is the composition of a venue type. The last column will be the standardized average price.

The dataset has 70 samples and more than 250 features. The number of features may vary for different runs due to FourSquare API may returns different recommended venues at different points in time.

Out[20]:


+
−
Avon Park
Bradenton
Sarasota
Port St. Lucie
North Port
West Palm Beach (33405): Palm Beach County
$719608.9
Cape Coral
Freeport
Naples
Big Cypress National Preserve
Everglades and Francis S. Taylor Wildlife Management Area
Everglades National Park
Leaflet

The number of features is much bigger than the number of samples. This will cause problem for the analysis process. Detail and counter-measurement will be discussed further in the next section.

III. Methodology:

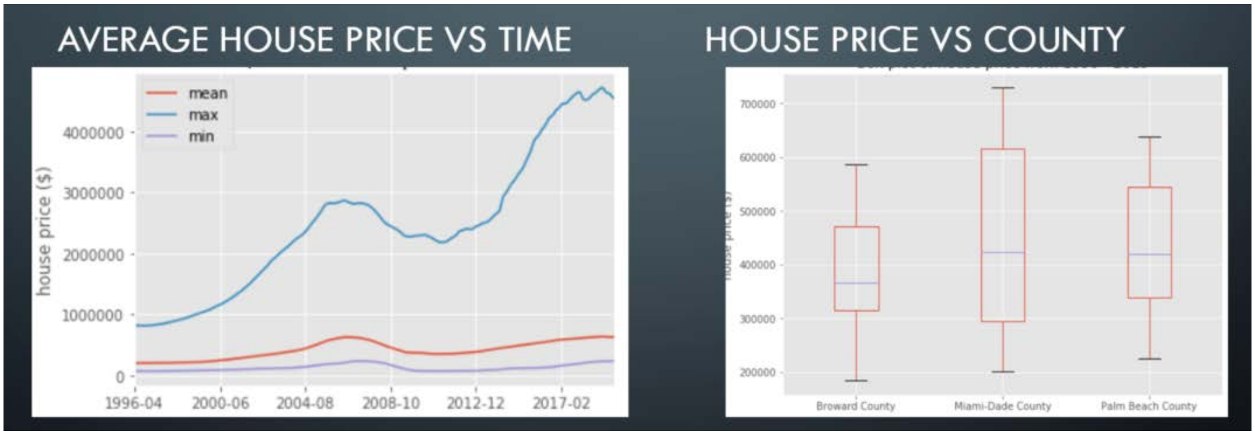
The assumption is that real estate price is dependent on the surrounding venue. Thus, regression techniques will be used to analyze the dataset. The regressors will be the composition of venue types. And the dependent variable will be standardized average prices.

At the end, a regression model will be obtained. Along with a coefficients list which describes how each venue type may be related to the increase or decrease of a neighborhood's real estate average price around the mean.

Python data science tools will be used to help analyze the data. Completed code can be found here: https://github.com/bjb96/Coursera_Capstone/blob/master/capstone-project-final.ipynb

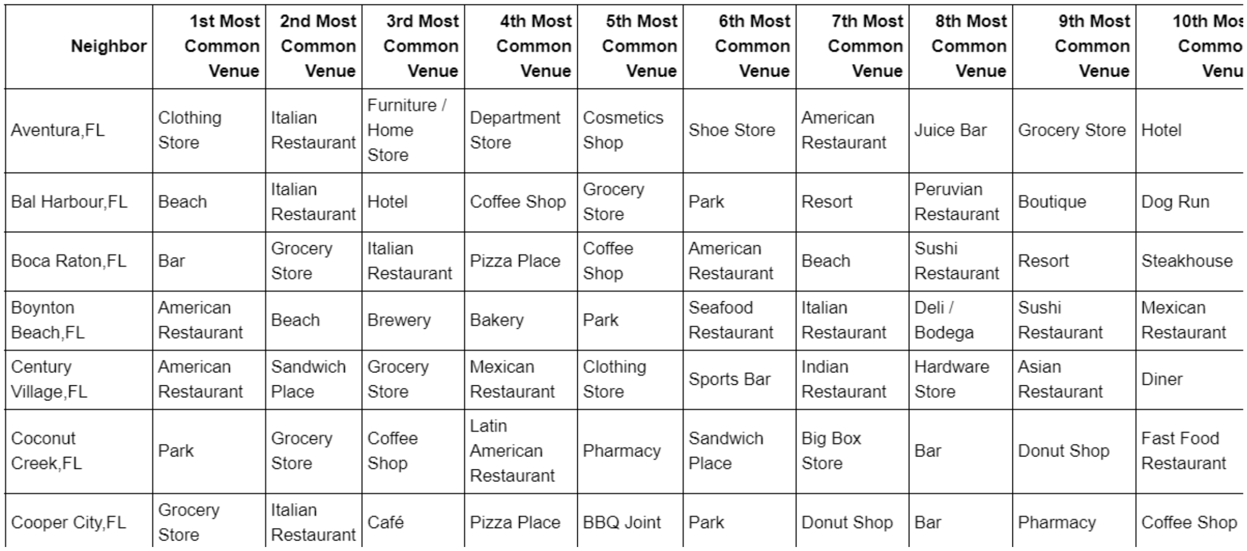
1. First insight using visualization:

In order to have a first insight of real estate average price between neighborhoods, there is no better way than visualization.



Above figure shows house price is significantly affected by time and location.

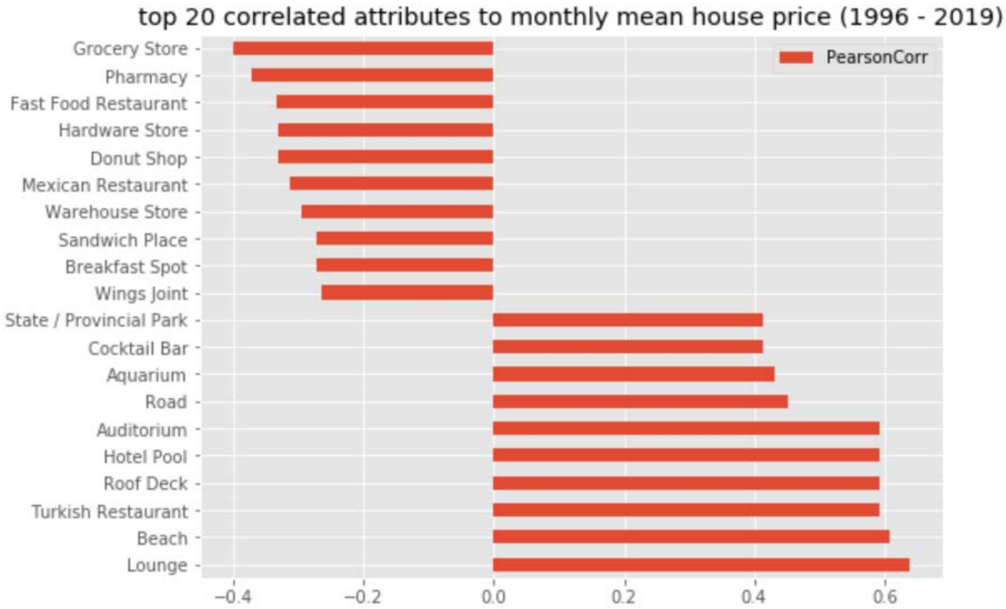
After aligning the venue composition data with neighborhood address, it is observed that the venue composition is very different among cities. An example is shown in below table.



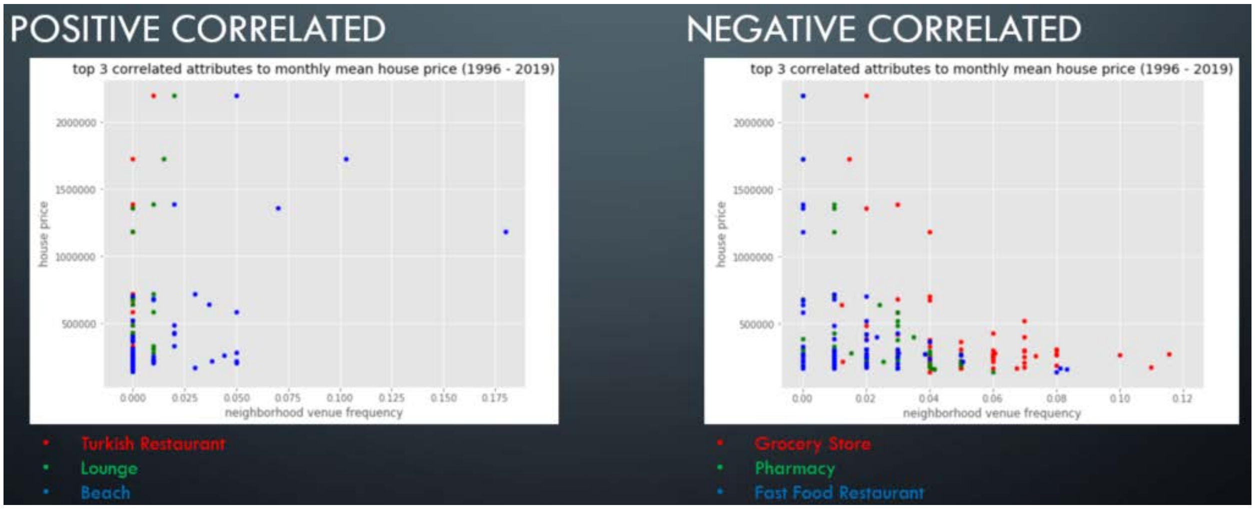
Neighbor	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Mos Commo Venu
Aventura,FL	Clothing Store	Italian Restaurant	Furniture / Home Store	Department Store	Cosmetics Shop	Shoe Store	American Restaurant	Juice Bar	Grocery Store	Hotel
Bal Harbour,FL	Beach	Italian Restaurant	Hotel	Coffee Shop	Grocery Store	Park	Resort	Peruvian Restaurant	Boutique	Dog Run
Boca Raton,FL	Bar	Grocery Store	Italian Restaurant	Pizza Place	Coffee Shop	American Restaurant	Beach	Sushi Restaurant	Resort	Steakhouse
Boynton Beach,FL	American Restaurant	Beach	Brewery	Bakery	Park	Seafood Restaurant	Italian Restaurant	Deli / Bodega	Sushi Restaurant	Mexican Restaurant
Century Village,FL	American Restaurant	Sandwich Place	Grocery Store	Mexican Restaurant	Clothing Store	Sports Bar	Indian Restaurant	Hardware Store	Asian Restaurant	Diner
Coconut Creek,FL	Park	Grocery Store	Coffee Shop	Latin American Restaurant	Pharmacy	Sandwich Place	Big Box Store	Bar	Donut Shop	Fast Food Restaurant
Cooper City,FL	Grocery Store	Italian Restaurant	Café	Pizza Place	BBQ Joint	Park	Donut Shop	Bar	Pharmacy	Coffee Shop

2. Correlation Analysis:

Pearson Correlation is used to reveal the relationships between composition of different venues and housing price of a neighborhood. The top 20 most correlated venues are shown below. Positive correlation means house price is higher when the venue category composition is also high. Negative correlation means house price is lower when the venue category composition is high



To understand the positive and negative correlations, scatter plots are used to show top 3 positive and negative correlated venues to the housing price as below.



POSITIVE CORRELATED

NEGATIVE CORRELATED

- Turkish Restaurant
- Lounge
- Beach

- Grocery Store
- Pharmacy
- Fast Food Restaurant

3. Feature Selection:

As mentioned before, there are more than 250 features (venues) for each neighborhood location, many of them are highly correlated between themselves. It is important to remove highly correlated features before conducting modeling to prevent overfitting.

Three methods are used to reduce the features, correlation, mutual information and F test. 11 features that have been selected by at least 2 of 3 methods are used in regression modeling.

4. Linear Regression:

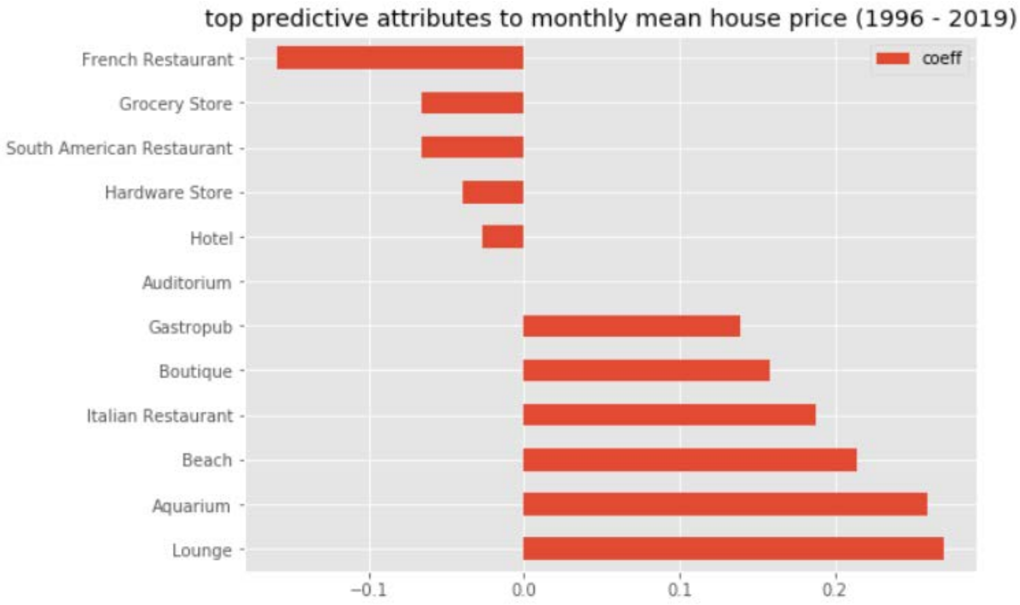
Linear Regression was chosen because it is a simple technique. And by using Sklearn library, implementing the model is quick and easy. Which is perfect to start the analyzing process.

The model will contain a list of coefficients corresponding to venue types. R2 score (or Coefficient of determination) and Mean Squared Error (MSE) will be used to see how well the model fit the data.

- The model’s R2 value is 48%
- The model’s MSE is 0.02

IV. Results:

Linear regression model is used to determine the predictive power of features. Regression model result shows the most influential venue composition for housing price. A figure is shown below,



V. Discussion:

Lounge, Aquarium and Beach have the most positive predictive power to housing price. This means the higher is their compositions in a neighborhood, the higher is the average housing price, and housing price is most sensitive to their composition variations

French Restaurants, Grocery Store, South American Restaurant and hardware store have the most negative predictive power. This means the higher is their compositions in a neighborhood, the lower is the average housing price, and housing price is most sensitive to their composition variations

VI. Conclusion:

Lower R2 means the venue composition is not the only housing price influential factor. As mentioned earlier, other factors such as school zones and crime rates also affect the housing price. A dataset incl. all the factors is expected to yield more accurate housing price prediction model.

Some notes on the project:

- It revealed housing price is mostly affected by the composition of Lounge, Aquarium, Beach, French Restaurants, Grocery Store, South American Restaurant and hardware Store within 5kM of a neighborhood
- It demonstrated the capability to conduct data analytical work using following tools and environment
  - Jupiter Lab / Python
  - IBM Cloud / Watson Studio
  - Github

Toward the person that went through this project, many thanks for the time and patient.