How “convenient” location can affect Washington D.C. neighborhood real estate price ?

Amin Oskouei
The University of Texas at Austin

Introduction

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

So, can the surrounding venus affect the price of a house? If so, what types of venues have the most affect, both positively and negatively?

Target Audience

- Potential buyers who can roughly estimate the value of a house based on the surrounding venues and the average price.
- Real estate makers and planners who can decide what kind of venues to put around their products to maximize selling price.
- Houses sellers who can optimize their advertisements.
- And of course, to this course's instructors and learners who will grade this project. Or to anyone who catch this shared on the social media showing that I can use Python data science tools.

Data Sources

The main data used for this project will be from two sources:

- The average Zillow home price index by neighborhoods in Washington, DC. Zillow
- The venues in each neighborhood. (FourSquare API)

Other supporting data:

- Coordinates (Geocoder Python)
- GeoJson (http://opendata.dc.gov/) and Github repository of DC maps (is available free at: https://github.com/benbalter/dc-maps

Data collection process

- Download zillow prices(csv file) for a list of Washington DC neighborhoods and their corresponding Zillow Home value index price.
- Find the geographic data of the neighborhoods. Both their center coordinates and their border.
- 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.

Final Data set look like this:

(73, 309)

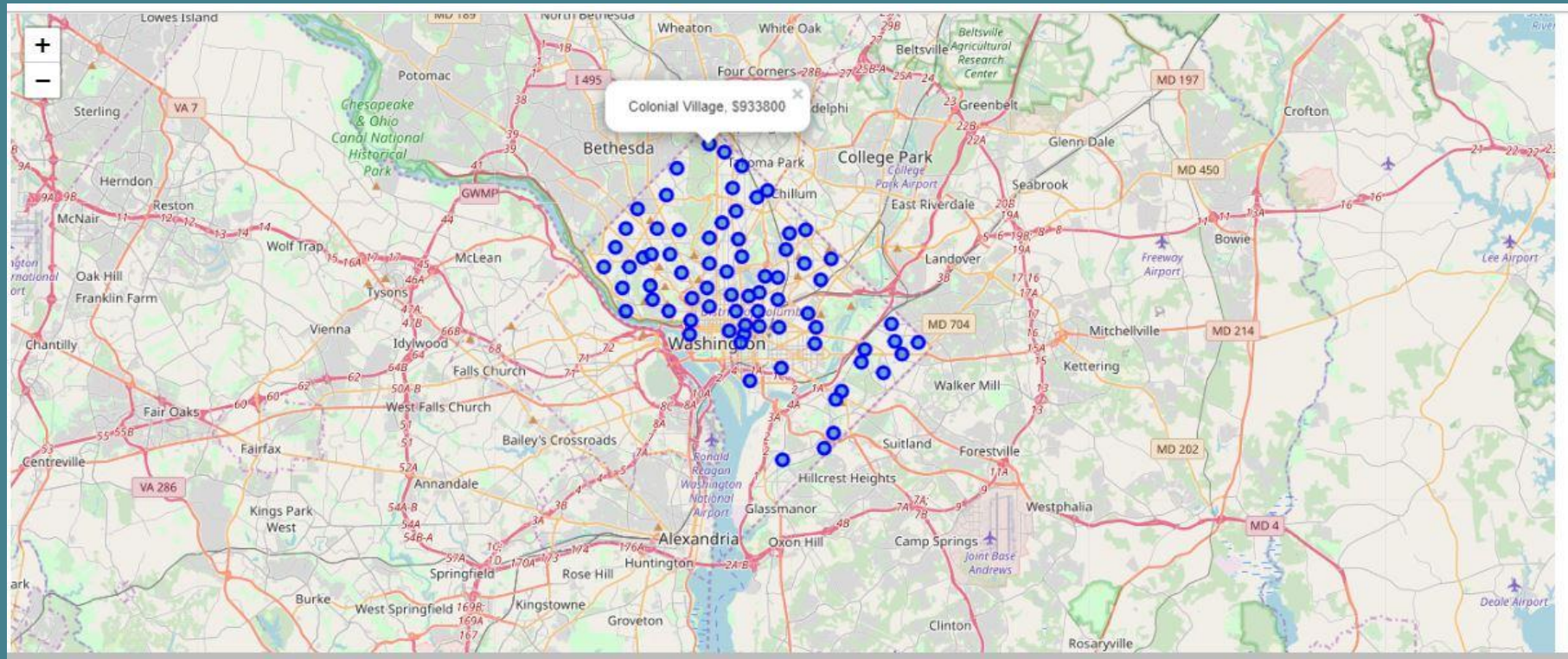
[220]:

	Neighborhood	Zoo Exhibit	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	American Restaurant	Arcade	Arepa Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	Arts & Entertainment	Asian Restaurant	Athletics & Sports
0	16th Street Heights	0	0	0	0	0	1	0	0	0	0	0	0	0	1
1	Adams Morgan	1	0	0	1	0	2	0	0	2	2	1	0	2	0
2	American University Park	0	0	0	0	1	2	0	0	0	1	1	0	1	0
3	Barnaby Woods	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	Benning	0	0	0	0	0	1	0	0	0	0	0	0	0	0

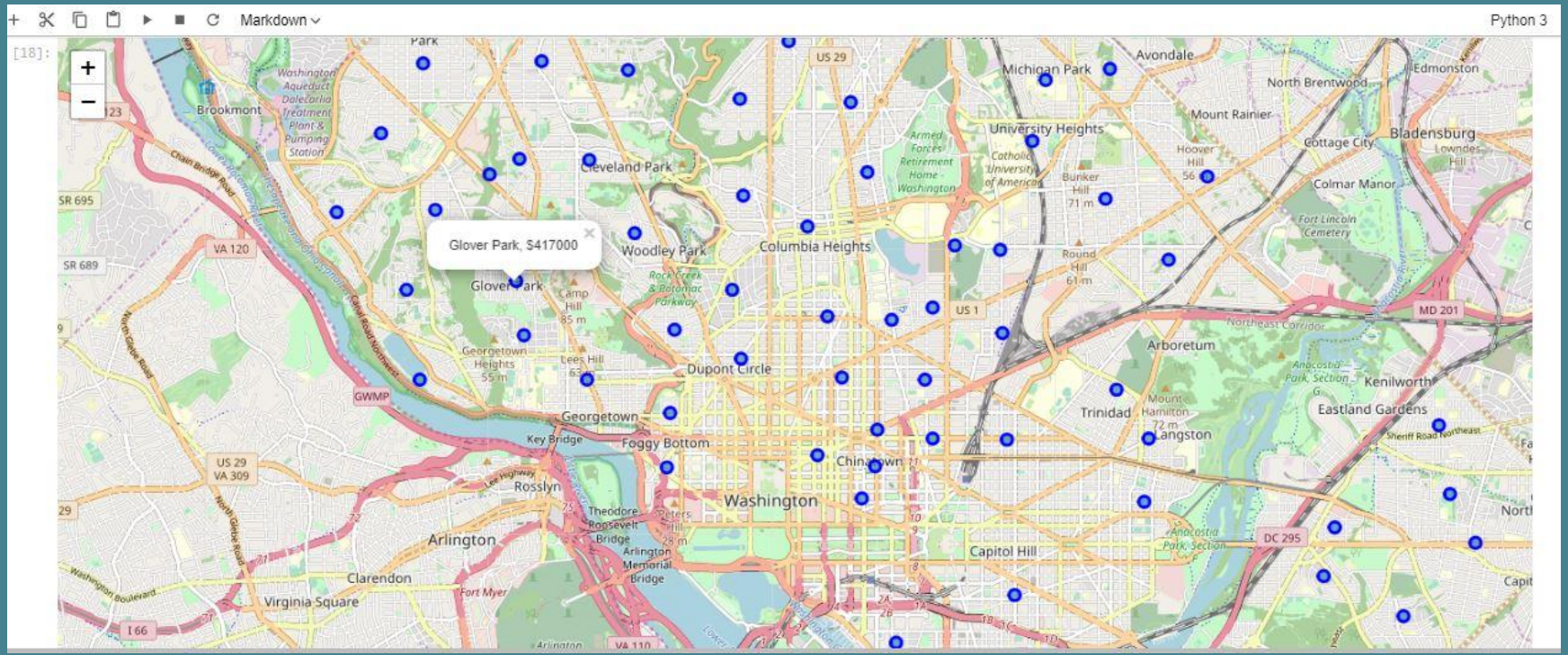
The output of the data collecting process will be a 2 dimensions dataframe

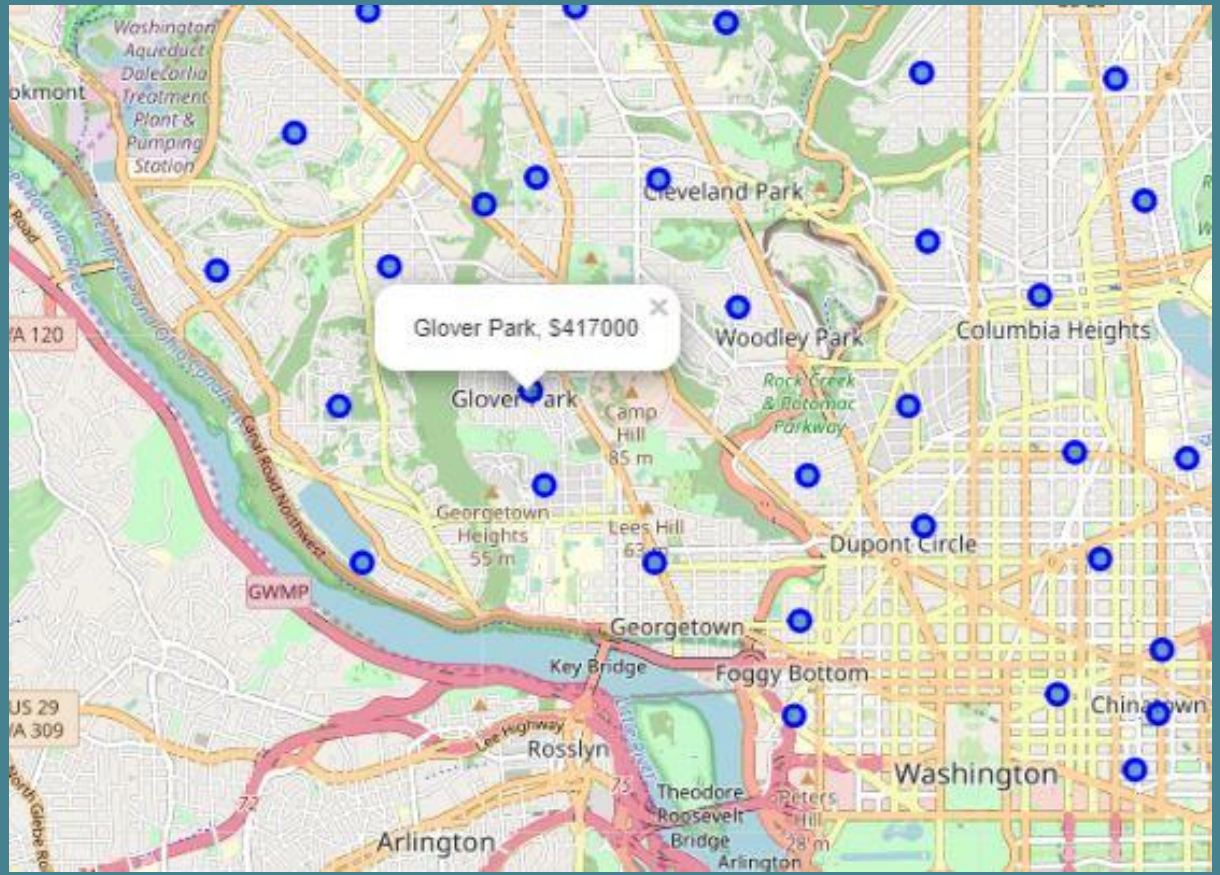
Data Visualization with Folium library in Python

Washington DC maps with neighborhood centroid markers



Washington DC maps with neighborhood centroid markers





We can see “Glover Park” have Zillow price index of $417K

Linear Regression:

```
print('Min coefs:', lreg.coef_[np.argsort(coef_abs)[:10]])
print('Venue types with least effect:', X.columns[np.argsort(coef_abs)[:10]].values)
```

```
R2-score: -0.47930992498036096
Mean Squared Error: 0.9324551521577173
Max positive coefs: [0.61929811 0.59880177 0.49479214 0.49479214 0.49324561 0.49285314
 0.47855573 0.475512   0.47155398 0.45666477]
Venue types with most postive effect: ['Historic Site' 'Skating Rink' 'Stadium' 'Roller Rink' 'Bus Line' 'Track'
 'General College & University' 'Botanical Garden' 'Souvenir Shop'
 'Baseball Stadium']
Max negative coefs: [-1.10310265 -0.56537993 -0.44723784 -0.44139399 -0.42661253 -0.42638938
 -0.42638938 -0.41100021 -0.35216761 -0.35216761]
Venue types with most negative effect: ['ATM' 'Harbor / Marina' 'Lake' 'Arcade' 'Herbs & Spices Store' 'Parking'
 'Fish & Chips Shop' 'Sculpture Garden' 'Storage Facility'
 'Video Game Store']
Min coefs: [ 0.          0.          0.          0.          0.          0.
  0.         -0.00010522  0.00073216 -0.0010244 ]
Venue types with least effect: ['Jewelry Store' 'Shoe Repair' 'Laundromat' 'Jazz Club'
 'Leather Goods Store' 'Lebanese Restaurant' 'Other Repair Shop'
 'Dessert Shop' 'Cantonese Restaurant' 'Building']
```

1. “Botanical garden” and “Historic site” both mean more tourist attractions. “Bus line” means ease of transportation. All of which usually increase the value of a location.
2. “ATM” and “Parking” and “Storage Facility” sure are nice to visit sometimes but may not be a suitable neighborhood for family with kids. “Lake” and “Sculpture Garden” usually located in the rural areas. The demand for such locations is usually low.
3. “Jazz Club”, “Shoe repair”, “Jewelry store”, all give value to a limited range of people.
4. “Gas Station” is available everywhere. These types of venue usually are not decision factor when considering a location.

Principal Component Regression (PCR):

[125]:
```
r2_max = scores_df['R2'].idxmax()
print("Best n:", r2_max, "R2 score:", scores_df['R2'][r2_max])

mse_min = scores_df['MSE'].idxmin()
print("Best n:", mse_min, "MSE:", scores_df['MSE'][mse_min])
```

```
Best n: 1 R2 score: -0.057128474003405705
Best n: 1 MSE: 0.6663410252521544
```

[126]:
```
# Use the best n_components parameter
lreg = LinearRegression()
X_train, X_test, y_train, y_test = train_test_split(X_pca[:,:r2_max], y, test_size=0.2, random_state=0)
model = lreg.fit(X_train, y_train)

# check the result
y_pred = lreg.predict(X_test)
r2 = r2_score(y_test, y_pred) # r2 score
mse = mean_squared_error(y_test, y_pred) # mse
print("R2 score:", r2)
print("MSE:", mse)
```

```
R2 score: -0.057128474003405705
MSE: 0.6663410252521544
```

We can see that R-squared improve but still negative.

The result seems to improved compared to just using simple Linear Regression.

Principal Component Regression (PCR):

[128]:
```
# Let's check which venue types effect the most and least
print('Max positive coefs:', pcr_coefs[np.argsort(-pcr_coefs)[:10]])
print('Venue types with most positive effect:', X.columns[np.argsort(-pcr_coefs)[:10]].values)
print('Max negative coefs:', pcr_coefs[np.argsort(pcr_coefs)[:10]])
print('Venue types with most negative effect:', X.columns[np.argsort(pcr_coefs)[:10]].values)
coef_abs = abs(pcr_coefs)
print('Min coefs:', pcr_coefs[np.argsort(coef_abs)[:10]])
print('Venue types with least effect:', X.columns[np.argsort(coef_abs)[:10]].values)
```

```
Max positive coefs: [0.00251286 0.00245616 0.00243678 0.00236941 0.002362   0.00236048
 0.00236034 0.0022442  0.00218617 0.00214282]
Venue types with most positive effect: ['Italian Restaurant' 'Cycle Studio' 'Salad Place' 'Ramen Restaurant'
 'New American Restaurant' 'Cocktail Bar' 'Ice Cream Shop'
 'Mediterranean Restaurant' 'Portuguese Restaurant' 'Theater']
Max negative coefs: [-0.00197313 -0.00167736 -0.00146918 -0.00131823 -0.00130322 -0.00127971
 -0.00127394 -0.0012362  -0.00120696 -0.00119246]
Venue types with most negative effect: ['Convenience Store' 'Pharmacy' 'Chinese Restaurant' 'Rental Car Location'
 'Video Store' 'Fast Food Restaurant' 'Mobile Phone Shop' 'Intersection'
 'Gas Station' 'Metro Station']
Min coefs: [-1.95104888e-05  2.85982565e-05  2.89792404e-05 -4.07235638e-05
 -4.07456446e-05  4.21252420e-05  4.70390463e-05  4.70390463e-05
  4.95027603e-05 -6.37871196e-05]
Venue types with least effect: ['Golf Course' 'Lounge' 'Waterfront' 'Track' 'Pool' 'Sports Bar'
 'Leather Goods Store' 'Jewelry Store' 'Spa' 'Supermarket']
```

Principal Component Regression (PCR):

1. This result shows that being close to “restaurant” like Italian, Romen, Cocktail bar and even “Ice cream shop” correlated with home price value (even though is not statistically significant)

2. Also, we can see “Rental Car location” , “Fast food”, “Gas station” , “Metro station” and “Pharmacy” have negative effect. Because this kind of store and location are not very attractive for high value neighborhood and its not economically worth it to open a gas station in a expensive lot.

Challenges:

1. Usually the needed data isn't publicly available.
2. When combining data from multiple sources, inconsistent can happen. And lots of efforts are required to check, research and change the data before merge.
3. For data obtained through API calls, different results are returned with different set of parameters and different point of time. Multiple trial and error runs are required to get the optimal result.
4. Even after the dataset has been constructed, lots of research and analysis are required to decide if the data should be kept as is or be transform by normalization or standardization.

Conclusion:

It's unfortunately that the analysis couldn't produce a precise model or showing any strong coefficient correlation for any venue type. But we can still get some meaningful and logical insights from the result.

- This project is done by a economist who only started self-studying Data Science for 12 months.
- The coefficients only show correlation, not causation. So, if your neighborhood average price is low, please don't go destroying the surrounding bars and food trucks. There might be another reason. For understanding differences between causation and correlation you can visit this page:

  https://www.georanker.com/correlation-vs-causality-differences-and-examples

References:

1. https://www.georanker.com/correlation-vs-causality-differences-and-examples
2. https://github.com/aminoskouei
3. http://opendata.dc.gov/
4. https://www.zillow.com/washington-dc/
5. https://en.wikipedia.org/wiki/Principal_component_regression