



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

04.01.2019

---

Amin Oskouei  
The University of Texas at Austin  
LinkedIn: <https://www.linkedin.com/in/amin-oskouei/>  
Github: <https://github.com/aminoskouei>

## Introduction

Founded on July 16, 1790, **Washington, DC** is unique among American cities because it was established by the Constitution of the United States to serve as the nation's capital. You can read the actual line at the National Archives. From its beginning, it has been embroiled in political maneuvering, sectional conflicts and issues of race, national identity, compromise and, of course, power.

Like many decisions in American history, the location of the new city was to be a compromise: Alexander Hamilton and northern states wanted the new federal government to assume Revolutionary War debts, and Thomas Jefferson and southern states who wanted the capital placed in a location friendly to slave-holding agricultural interests.

It remains a vibrant and culturally diverse city today. The city is rich with international cultures, African American heritage and culture and it's also one of America's most gay-friendly cities. In fact, DC recognized same-sex marriage in 2010, before the Supreme Court, nearby, ruled that it was a right in 2015.

After more than 200 years as the nation's capital, Washington has developed as a complex and layered city, with a distinctive character: both a town for locals, an international center of power and an amazing place to visit.

Today, you can have the best of both worlds by delving into the nation's past with a visit on the National Mall and museums or adventuring into very modern, exciting neighborhoods. No wonder the real estate market in Washington, DC is in high Demand due to many reasons like job market and government agencies and plenty of educational institution. It's where a House in North capitol street is listed for sale around [\\$1.75 million](#) (\$619/ft<sup>2</sup>).

But that is just an outlier example. A quick search can show us the real estate price can vary by a large margin from neighborhoods to neighborhoods. For example, a 1-bedrooms condo in Adam morgan, can cost [\\$400K](#); while a 1-bedrooms in benning, it's only [\\$100 thousands](#).

So what aspects of a neighborhood that can affect the price of real estates to such extent? One hypothesis is that the surrounding venues can be a decision factor. Surely anyone, who has attempted to find an accommodation for rent or buy, has seen advertisements such as: This condo is located near the subway station, malls, supermarkets, Public Transportation and dinners, etc. And it's likely that the price will be higher than others with locations not as "convenient".

Can the venues surrounding an accommodation affect its price? And what kind of venues can affect the most? How much of the price affected by location of a Apt or House?

## Goals

This project will try to explore the neighborhoods of **Washington, DC** to see:

- if the surrounding venues can affect the price of real estates?
- what kind of surrounding venues, and to what extend, can effect the price?
- if we can use the surrounding venue to estimate the value of an accommodation over the average price of one area? And to what degree of confidence?

The result can be useful for home buyers, who can roughly estimate the value of a target house over the average. Or to planners, who can decide which venues to place around their product, so that the price is maximized. Or to just anyone that curios about Washington D.C. and real estate.

The target audience for this report are:

- 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>)

Let me explain more about zillow home value index:

### What's the Zillow Home Value Index?

Before we tackle the Zillow Home Value Index, be sure to learn about the Zestimate home valuation, since this is the building block for the Zillow Home Value Index. A Zestimate is Zillow's estimate of the current market value for a home. We have tens of millions of



Zestimates – one for most homes. Our data is refreshed regularly to reflect real estate transactions that could affect you – even if you’re not buying or selling a house.

### **what's the Zillow Home Value Index?**

The Zillow Home Value Index is the median Zestimate valuation for a given geographic area on a given day.

Can you give me another example?

Sure. Let's take Seattle. On Aug. 19, 2009, the Zillow Home Value Index for single-family homes in Seattle was \$373,714, which means half the homes have values less than \$373,714 and half have values greater than \$373,714.

## **Data collection process**

Data collection process:

- The average price will be scrapped from the Zillow website.
- For each neighborhood, use Geojson file of neighborhood to get its coordinate.
- For each neighborhood's coordinate, call FourSquare API to get the surrounding venues.
- Count the occurrences of each venue type and attach that information to each neighborhood.

The process of collecting and clean data:

- 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.

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

- Each row represents a neighborhood.
- Each column will be the count of one type of venue in that neighborhood.
- The last column will be the Zillow home value index (for all type of house) of that neighborhood.

The result dataset is a 2 dimensions data frame (Figure 1):

(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

Figure 1. The final Data set using for regression analysis

The dataset has 73 samples and more than 309 features. The number of features may vary for different runs due to FourSquare API may returns different recommended venues at different points in time.

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.

## 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 occurrences 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/aminoskouei/IBM-Data-Science/blob/master/DC\\_Capstone\\_week5\(final%20code\).ipynb](https://github.com/aminoskouei/IBM-Data-Science/blob/master/DC_Capstone_week5(final%20code).ipynb)



## 1. First insight using visualization:

In order to have a first insight of Washington D.C. real estate Zillow price index between neighborhoods, there is no better way than visualization.

The medium chosen is centroid street open neighborhood map, which has markers that show the centroid location of each neighborhood on the map with price labels. So, we can easily explore each neighborhood and find out the Zillow price index for that neighborhood.

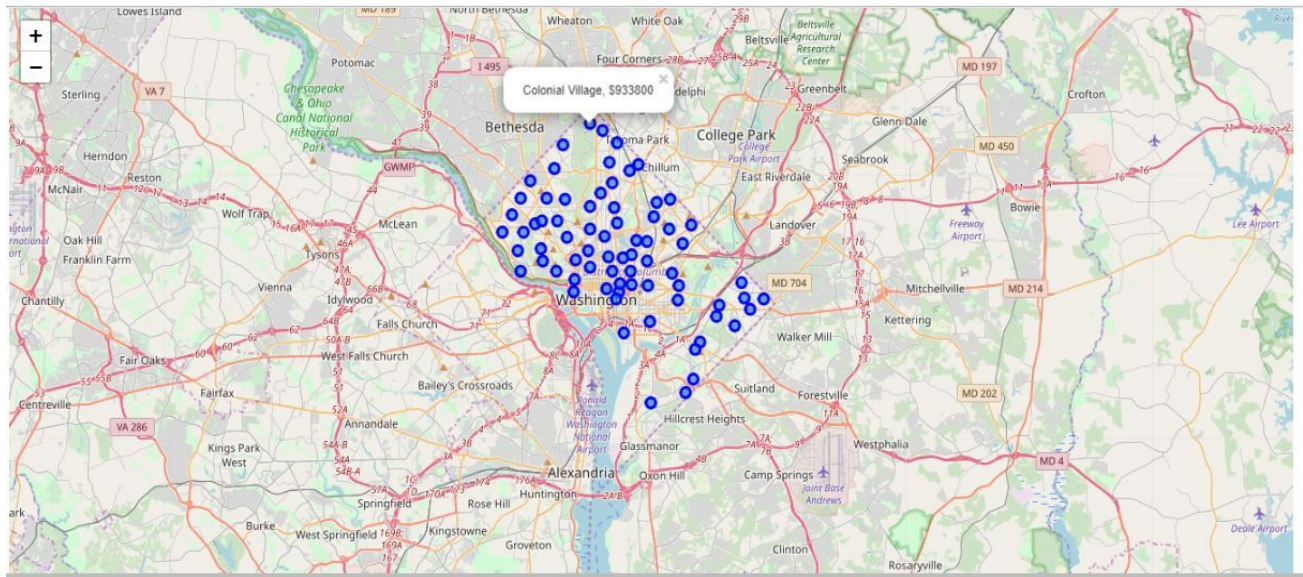


Figure 2. Show the map of Washington D.C. with neighborhood centroid point (Blue point).

Let's look closer to the map to see how's the label look like:

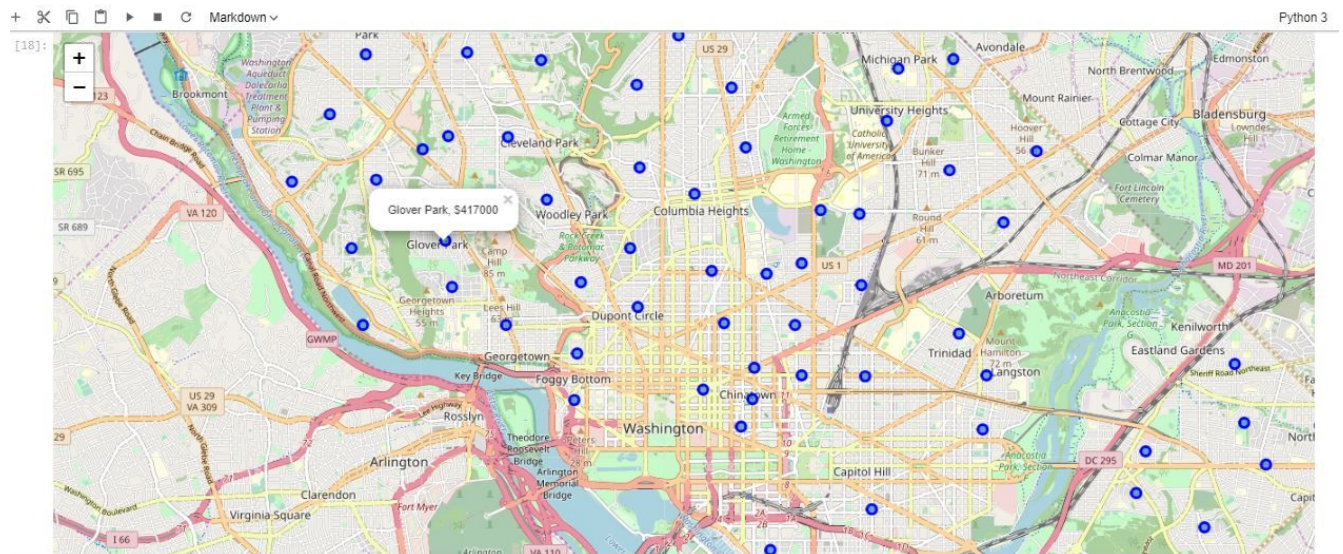


Figure 3. Folium Map of Washington D.C neighborhoods

We zoom into “Glover Park” neighborhood and we can see the Zillow home value index for that neighborhood is \$417K.

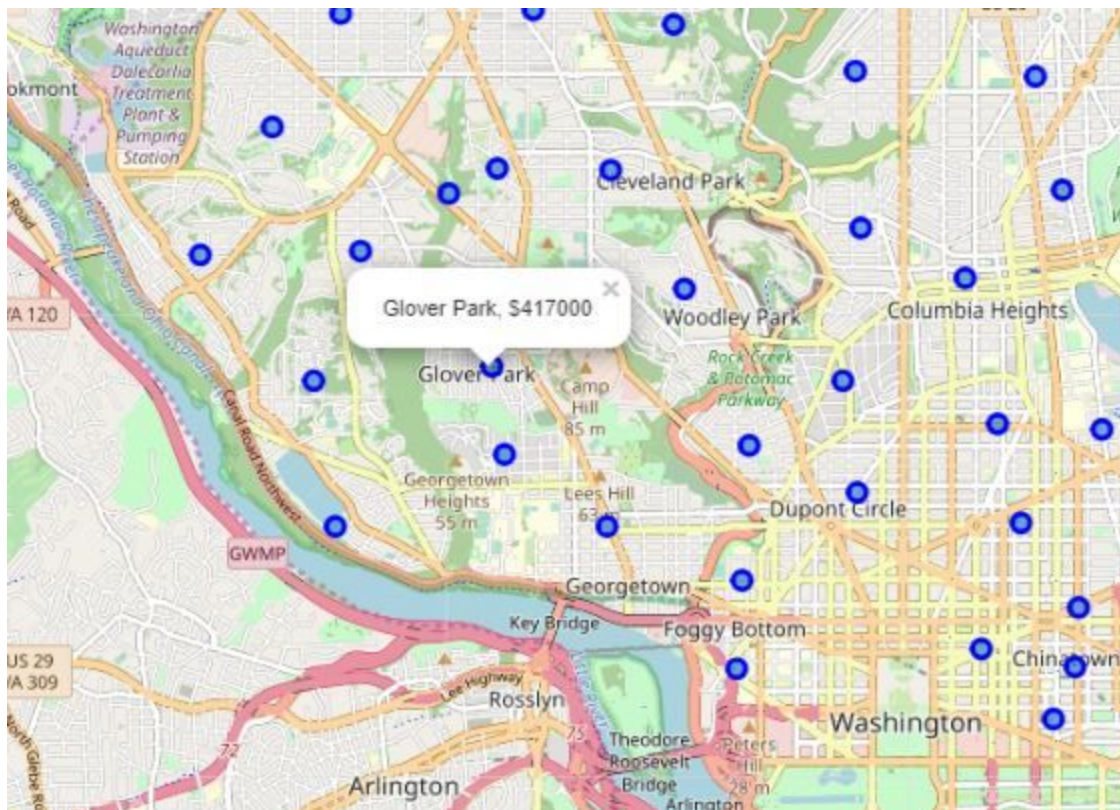


Figure 4. Glover Park neighborhood on the Map



## 2. 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 result (Figure 5) doesn't seem very promising. R2 score is small, which means the model may not be suitable for the data. We have problems in this model's because we need to control regression for home characteristic to get proper result but because the goal of this project is more about using data science tools and technique we ignore this issues.

```
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 positive 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']

Figure 5. The coefficient of linear regression

But on the bright side, the coefficient list shows some interest and logical information:

- “Botanical garden” and “Historic site” both mean more tourist attraction. “Bus line” means ease of transportation. All of which usually increase the value of a location.
- “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.
- “Jazz Club”, “Shoe repair”, “Jewelry store”, all give value to a limited range of people.
- “Gas Station” is available everywhere. These types of venue usually are not decision factor when considering a location.



Back to the model, what seems to be the problem? And what are the possible solutions?

Looking back further to the dataset, its dimensions sizes is clearly unbalanced, only 73 samples, and more than 300 features. Logical steps to take are either collecting more samples or trying to reduce the number of features.

But since there are no other public source available, increasing sample size is not possible at the moment. So, decreasing features is the only option for now.

And that's why Principal Component Regression is chosen to analyze the dataset in the next part.

### 3. Principal Component Regression (PCR):

PCR can be explained simply as the combination of Principal Component Analysis (PCA) with Linear Regression.

PCR employs the power of PCA, which can convert a set of values of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. As the result, the number of features is reduced while keeping most of the characteristic of the dataset.

```
[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
```

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

Figure 6. PCR result

Then PCR use Linear Regression on the converted set to return a coefficient list, just like in normal Regression techniques.

Again, R2 score and MSE are used to see how well the model fit the dataset.

The result is promising as it shows improvement over the simple Linear Regression.

As for the coefficient list, the size has been reduced after performing PCA. So, a dot product with eigenvectors is needed to get it back to the original features size.

```
[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']
```

Figure 7. The final result for regression

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.

## Results

Even though the scores seem to be improved after applying a more sophisticated method, the model is still not suitable for the dataset. Thus, it can't be used to precisely predict a neighborhood average price.

Explanations for the poor model can be:

- The real estate price is hard to predict.
- The data is incomplete (small sample size, missing deciding factors).
- The machine learning techniques are chosen or applied poorly.

But again, on the bright side, the insight, gotten from observing the analysis results, seems consistent and logical. And the insight is business venues that can serve the needs of most normal people usually situated in pricey neighborhoods.

## Discussion

The real challenge is constructing the dataset:

- Usually the needed data isn't publicly available.
- When combining data from multiple sources, inconsistent can happen. And lots of efforts are required to check, research and change the data before merge.
- 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.
- 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.

It can be considered the most important process in the whole data science pipeline. Which can affect the most on the result.

On the other hand, choosing the suitable technique to construct the model is also a worthwhile process. As this report shows that, by applying a different method, the result can be improved.

## 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.

Doing this project helps practicing every topic in the specialization, and thus, equipping learners with Data Science methodology and tools using Python libraries. Also doing a real project certainly helps one learns so much more outside the curriculum, as well as realizes what more to research into after completing the program. And as this report shows, there are surely a lot of things to dig into.

Some notes on the analysis 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>

Toward the person that went through this project, thanks for your time and patience.

## 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](https://en.wikipedia.org/wiki/Principal_component_regression)