IBM – Coursera

Data Science Specialization

Capstone project - Final report

# Correlation between the Maryland Counties rental value and its surrounding venues

By Noelia Meddens

# Table of content:

# I. Introduction:

This report is for the final task of the Data Science Specialization Certification created by IBM. The problem and the analysis approach must to be done using the Foursquare location data to explore or compare neighborhoods or cities.

The principal goal will be exploring the Maryland counties to find the correlation between the house rental value and its surrounding venues.

The idea comes from the socioeconomic study of the state, being this only first stage. It’s common that minority groups, immigration and low employment are related with prices of life and services, such as community institutions and international restaurants.

So, can the surrounding venues affect the rental value? If so, what types of venues have the most affect, both positively and negatively?

The target audience for this report are:

- Government institutions
- NGO's
- And anyone interested about Data science and/or help the poor cities in anyway possible.

## II. Data description:

Maryland is one of the counties that includes a variety of cities, including some of the most poor and some of the biggest differences in racial groups differences. Other reasons are the data available, as I found many sources with all the socioeconomic data of the area.

The dataset will be composed from the following two main sources:

- OpenDataSoft where you can find worldwide information available to download. https://public.opendatasoft.com/
- FourSquare API which provides the surrounding venues of a given coordinates.

The process of collecting and clean data:

- Download the necessary data in .csv format and import to watson.
- Merge, filter and clean the data.
- For each city, 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 city. Then apply one hot encoding to turn each venue type into a column with their occurrence as the value.
- Standardize the average value by removing the mean and scaling to unit variance.

The final dataset is a 2 dimensions data frame :

- Each row represents a city.
- Each column, except the last one, is the occurrence of a venue type. The last column will be the standardized average rental value.

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.

# 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 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/Swintha/Coursera_Capstone/blob/master/Maryland_Capstone%20IBM%20Data%20Science(1).7z

1. First insight using visualization:

In order to have a first insight of Maryland rental value between Cities, there is no better way than visualization.

The medium chosen is Choropleth map, which uses differences in shading or coloring to indicate the rental value within predefined areas. It is ideal for showing the differences across the Maryland State map.

The map (Figure 2) shows high rental value in the areas located nearby Washinton.

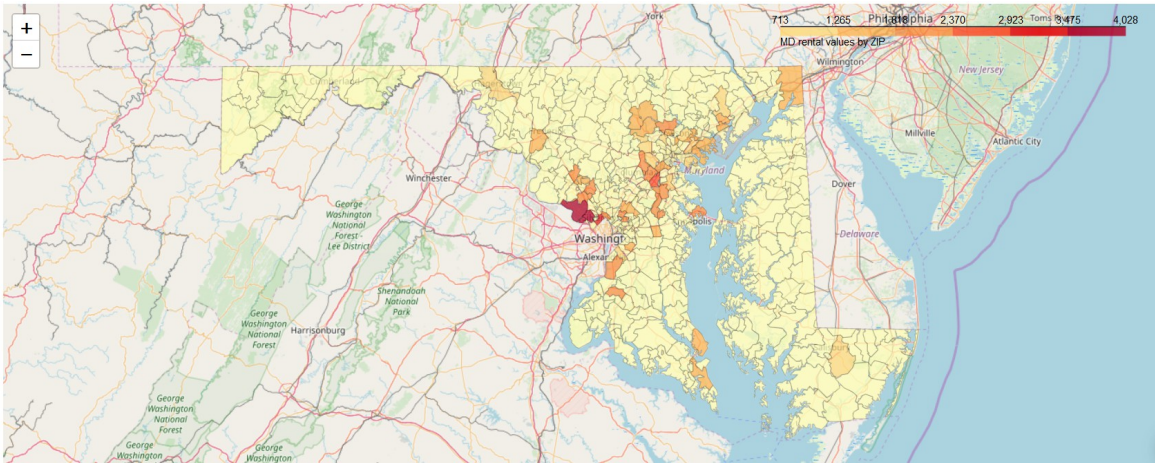

*Figure 1 - Maryland State Rental Value*

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 3) doesn't seem very promising. R2 score is small, which means the model may not be suitable for the data.

```
R2-score: -8.858501653124736e+22
Mean Squared Error: 3.5478193926057434e+28
Max positive coefs: [2.60188552e+14 1.82490619e+14 1.82490619e+14 1.76436648e+14
 1.16709704e+14 1.07729240e+14 9.04176516e+13 8.80303916e+13
 8.15406526e+13 8.15406526e+13]
Venue types with most postive effect: ['Beach' 'Filipino Restaurant' 'Event Space' 'Big Box Store'
 'Botanical Garden' 'Bagel Shop' 'Performing Arts Venue' 'Comic Shop'
 'Racetrack' 'Convention Center']
Max negative coefs: [-9.15714527e+14 -8.64842815e+14 -2.95739461e+14 -2.64566071e+14
 -2.51312693e+14 -2.49129947e+14 -2.44498878e+14 -2.01920014e+14
 -1.51833461e+14 -1.35982151e+14]
Venue types with most negative effect: ['Accessories Store' 'Argentinian Restaurant' 'Burrito Place' 'Arcade'
 'Cemetery' 'Afghan Restaurant' 'Social Club' 'Boutique'
 'Caribbean Restaurant' 'Aquarium']
Min coefs: [ 0.00000000e+00  0.00000000e+00  0.00000000e+00 -1.28754653e+11
 -2.12031627e+11  5.22595289e+11  6.06195185e+11  6.77683888e+11
 -7.37051386e+11 -1.15543349e+12]
Venue types with least effect: ['Comedy Club' 'Water Park' 'Gun Range' 'Concert Hall' 'Hotel'
 'History Museum' 'Church' 'Pizza Place' 'Food Truck' 'Convenience Store']
```

*Figure 2 - Linear Regression result*

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

- The venues with related with high Social Status increase the rental value
- The decrease of the rental value is related with business for people with low incomes and social services.

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

Looking back further to the dataset, the quality is not the best as we are missing information due the public and free information. There is not quality data available to perform a better study.

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. (Wikipedia, n.d.)

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.

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.

```
R2 score: 0.15694132999421595
MSE: 337643.99620514584
```

*Figure 3 - PCR scores*

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.

```
Max positive coefs: [72.07705657 66.4902946  59.02128296 58.74543608 58.34704262 54.62555373
 53.21449893 44.78690155 44.10999884 43.02730719]
Venue types with most positive effect: ['Eastern European Restaurant' 'Wine Bar' 'Performing Arts Venue' 'Hotel'
 'Church' 'Beer Bar' 'Korean Restaurant' 'Shopping Plaza'
 'Falafel Restaurant' 'Waterfront']
Max negative coefs: [-86.64883675 -64.45242376 -64.45242376 -48.3593521  -45.95852
 -45.18471879 -41.60663723 -40.9076193  -40.9076193  -39.4092455 ]
Venue types with most negative effect: ['Gift Shop' 'Dog Run' 'Other Great Outdoors' 'Martial Arts Dojo'
 'Vietnamese Restaurant' 'Movie Theater' 'Steakhouse' 'Mountain'
 'Outlet Mall' 'Neighborhood']
Min coefs: [ 0.20033834  0.37668992  0.37668992  0.43150877  0.47814857 -0.57802474
  0.61818218  0.77232925 -0.85471307 -0.85471307]
Venue types with least effect: ['Smoothie Shop' 'Water Park' 'Gun Range' 'Breakfast Spot' 'Wings Joint'
 'Comedy Club' 'Electronics Store' 'Deli / Bodega'
 'College Basketball Court' 'Burrito Place']
```

*Figure 4 - Coefficient list in original size*

The insight is still consistent compared to the Linear Regression's.

# IV. Results:

Even though the scores seem to be improved after applying a more sophisticate method, the model is still not suitable for the dataset. Thus, it can't be used to precisely predict the cities rental value.

Explanations for the poor model can be:

- 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 cities with more minority social groups (Unemployment, poverty, immigration,etc.) are more likely to have a low rental value.

# V. 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.

# VI. 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.

# References:

https://www.geeksforgeeks.org/types-of-regression-techniques/

# Table of Figures: