CORRELATION BETWEEN NEIGHBORHOOD REAL ESTATE PRICE AND ITS SURROUNDING VENUES

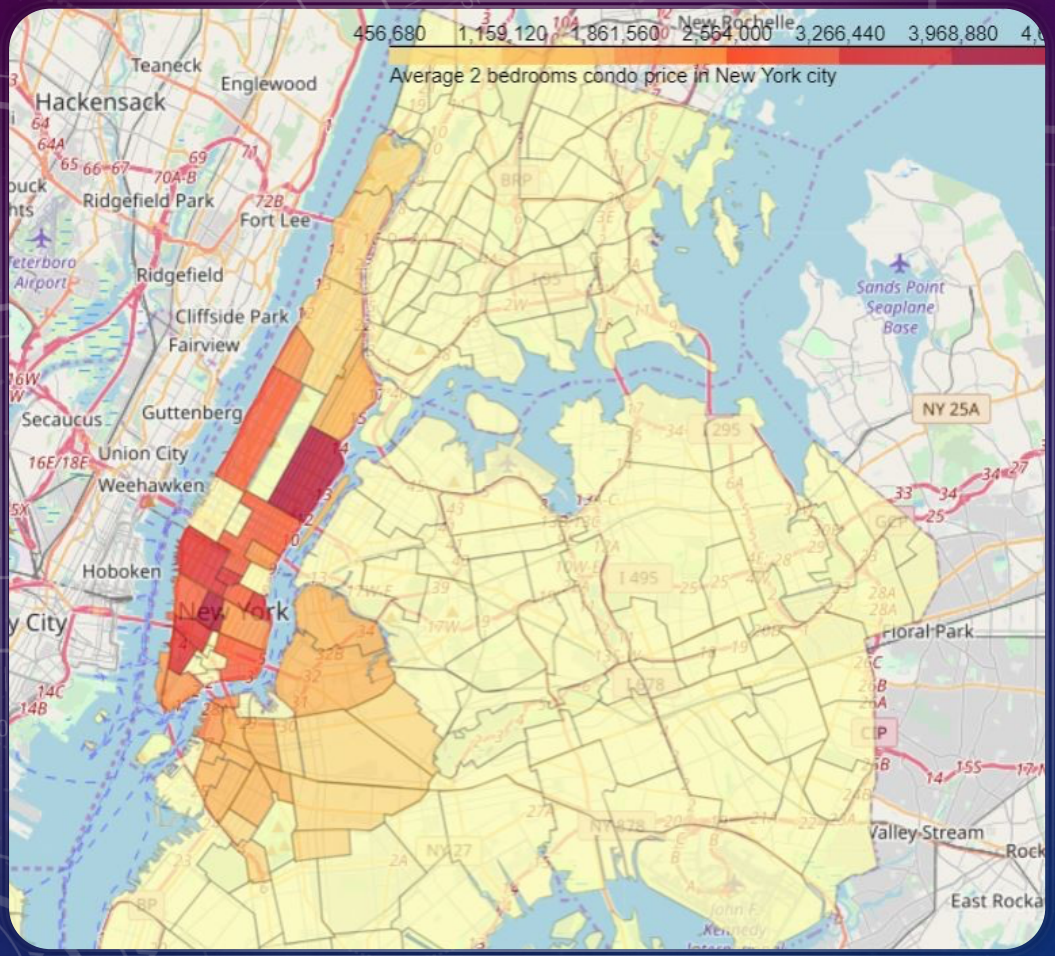
PROJECT INTRODUCTION:

- This project is the capstone for the 4-courses IBM Applied Data Science Specialization on Coursera. The requirement is leveraging the Foursquare location data to explore or compare neighbourhoods or cities of your choice or to come up with a problem that you can use the Foursquare location data to solve.
- The chosen topic is the correlation between the real estate value and its surrounding venues.
- The idea comes from the process of, any family searching a home to stay after moving to another city. It is common that the owners or agents advertise their properties are closed to some kinds of venues like supermarkets, restaurants, public transport, hospital, school or coffee shops, etc.; showing the “convenience” of the location in order to raise their house’s value to sale or rent.

DATA ACQUISITION AND CLEANING:

	Neighborhood	AvgPrice	Latitude	Longitude
0	Bedford-Stuyvesant	750000	40.687232	-73.941785
1	Boerum Hill	1.69e+06	40.685683	-73.983748
2	Brooklyn Heights	2.15e+06	40.695864	-73.993782
3	Bushwick	967000	40.698116	-73.925258
4	Carroll Gardens	1.51e+06	40.680540	-73.994654

- **1. Scrap CityRealty website for neighborhoods average prices:**
- URL: https://www.cityrealty.com/nyc/market-insight/features/get-to-know/average-nyc-condo-prices-neighborhood-june-2018/18804
- **2. Get the neighborhoods coordinate:**
- Free geodata is available free at: https://geo.nyu.edu/catalog/nyu_2451_34572



NEW YORK CITY NEIGHBOURHOODS IN MANHATTAN AND BROOKLYN WERE CHOSEN AS THE OBSERVING TARGET.

ANALYSIS:

1. CHECK FOR CORRELATION BETWEEN *OCCURRENCE OF SURROUNDING VENUES* WITH *REAL ESTATE AVERAGE PRICE*:

```
R2-score: -0.022911025942468966
Mean Squared Error: 0.35802873946081304
Max positive coefs: [0.3777725  0.30044348 0.28204759 0.28204759 0.27855932 0.24938489
 0.24938489 0.24938489 0.24938489 0.22125059]
Venue types with most postive effect: ['General Entertainment' 'Other Nightlife' 'Cafeteria' 'Buffet'
 'Colombian Restaurant' 'Jewish Restaurant' 'Train Station'
 'Persian Restaurant' 'Resort' 'Dumpling Restaurant']
Max negative coefs: [-0.25207753 -0.22175376 -0.22175376 -0.21377297 -0.21377297 -0.21377297
 -0.20118393 -0.18698446 -0.17897163 -0.17897163]
Venue types with most negative effect: ['Board Shop' 'Flea Market' 'Golf Driving Range' 'Street Food Gathering'
 'Print Shop' 'Other Repair Shop' 'Drugstore' 'Street Art'
 'Gluten-free Restaurant' 'Sports Club']
Min coefs: [0. 0. 0. 0. 0. 0. 0. 0. 0. 0.]
Venue types with least effect: ['Gym Pool' 'Indoor Play Area' 'Bridge' 'TV Station' 'Food Stand'
 'Shipping Store' 'Cemetery' 'Gas Station' 'Hookah Bar' 'Factory']
```

- THE RESULT DOESN'T LOOK PROMISING:
- THE R2 SCORE IS SMALL.
- THERE ARE NO REALLY STRONG COEFFICIENT CORRELATIONS.

2. APPLYING PCR FOR BETTER RESULT:

```
Max positive coefs: [0.06348595 0.06289923 0.05825005 0.05691995 0.05108069 0.05090563
 0.04779149 0.0463552  0.0463552  0.04494847]
Venue types with most positive effect: ['Dumpling Restaurant' 'Design Studio' 'Pilates Studio' 'Library'
 'Korean Restaurant' 'Colombian Restaurant'
 'Southern / Soul Food Restaurant' 'Buffet' 'Cafeteria' 'Sushi Restaurant']
Max negative coefs: [-0.05611648 -0.0484223  -0.04423582 -0.04191699 -0.04188961 -0.04080127
 -0.03924051 -0.03762934 -0.03722191 -0.03624661]
Venue types with most negative effect: ['Market' 'Trail' 'Food & Drink Shop' 'Tapas Restaurant' 'Lingerie Store'
 'Garden' 'New American Restaurant' 'Coffee Shop' 'Street Art'
 'Peruvian Restaurant']
Min coefs: [-9.29409627e-06 -2.05033683e-05 -1.60035383e-04 -1.60035383e-04
 -1.88320212e-04 -1.88320212e-04 -1.88320212e-04  1.91183662e-04
 -1.99128786e-04 -2.35729534e-04]
Venue types with least effect: ['Vape Store' 'State / Provincial Park' 'Molecular Gastronomy Restaurant'
 'Volleyball Court' 'Laser Tag' 'Factory' 'Boat or Ferry' 'Food Court'
 'Indoor Play Area' 'Bookstore']
```

```
R2 score: 0.454460324852
MSE: 0.190944155714
```

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

RESULT:

- Based on the assumption that the price of a real estate is dependent on its surrounding venues. Regression techniques were used to get the coefficient correlation between each venue type and the price. And at the end, producing a model to predict how higher or lower a neighbourhoods price compared to the mean, based on the occurrence of its surrounding venue types.
- First, Simple Linear Regression was used to see how the approach would perform. Then a more sophisticated method, Principal Component Regression (PCR), was applied to improve the result.
- Unfortunately, the end result isn't very promising. With the R2 score (or Coefficient of determination) of 0.45, the model isn't really fit to the data set; and thus, can't be used for further predicting the real estate price.

BUT ON THE BRIGHT SIDE, INTERPRETING THE COEFFICIENT LIST GAVE AN INSIGHT THAT SEEMS TO BE LOGICAL:

- Venue types with most positive effect: ['Dumpling Restaurant' 'Pilates Studio' 'Design Studio' 'Pie Shop'
- 'Southern / Soul Food Restaurant' 'Library' 'Sushi Restaurant' 'Resort'
- 'Korean Restaurant' 'Buffet']
- Venue types with most negative effect: ['Market' 'Lingerie Store' 'Gay Bar' 'Kosher Restaurant' 'Optical Shop'
- 'Food' 'Food Truck' 'Wine Bar' 'Food & Drink Shop' 'Climbing Gym']
- Venue types with least effect: ['Christmas Market' 'TV Station' 'Cemetery' 'Event Space'
- 'Indoor Play Area' 'Modern European Restaurant' 'Mini Golf'
- 'Volleyball Court' 'Molecular Gastronomy Restaurant' 'Community Center']

- Venue types like "Studios" or fancy "Eateries" usually located in busy areas, where there are lots and lots of people seeking their services. And that's usually where people, who love the bustling atmosphere of New York city, would love to live nearby. High demand equals high price.
- On the other hand, venue types like "TV station", "Cemetery", "Mini Golf", etc. usually are not important factors to most people when looking into a neighborhood. So, they don't have much effect to the price.

CONCLUSION AND FUTURE DIRECTIONS:

- First, about the data. Real estate prices are not usually available to the public. So, collecting a large set is impossible without connections with some real estate agencies. In this project, there are only 50 samples, but with more than 300 features. Since collecting more samples is not possible at the moment, PCR was chosen to solve the problem by reducing the features size before applying regression.
- Second, about the analysis process and conclusion. With no formal academic background in statistics and mathematics, the tools and methods might not be used with their optimal configuration. And the insight might not be drawn out fully, or even worse not correct at all. Further study in statistical inference and multivariate statistical analysis after this program is a must.