**I. Description of the Project:**

**1. Introduction:**

Who ever has come to New York will be enchanted by her beauty and atmosphere, with vibrant life 24/7, never-ending to-do list, booming business at every corners, and by the expertly-planned transport system which can take you everywhere in the city.  
New York is famous worldwide as the finance center and the "tourist must go to" city in North America. No wonder the real estate market there is one of the priciest. It's where a penthouse at the new Peter Marino-designed condo The Getty, just steps from the High Line, is sold for $59 million (\\$5,826/ft2).  
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 2-bedrooms condo in Central Park West, Upper West Side can cost \$4.91 millions on average; while in Inwood, Upper Manhattan, just 30 minutes aways, it's only \\$498 thousands.

So what aspects of a neighborhood that can effect the price of real estates to such extend? 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, 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 effect its price? And what kind of venues can effect the most? And by what weight?

**2. The question to solve:**

This project will try to explore the neighborhoods of New York city to see:

* if the surrounding venues can effect 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 any normal person

**II. Description of the data:**

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

* The average price by neighborhoods in New York city. ([CityRealty](https://www.cityrealty.com/nyc/market-insight/features/get-to-know/average-nyc-condo-prices-neighborhood-june-2018/18804))
* The venues in each neighborhood. ([FourSquare API](https://developer.foursquare.com/))

Other supporting data:

* Coordinates ([Geocoder Python](https://geocoder.readthedocs.io/index.html))
* GeoJson ([http://data.beta.nyc](http://data.beta.nyc/dataset/0ff93d2d-90ba-457c-9f7e-39e47bf2ac5f/resource/35dd04fb-81b3-479b-a074-a27a37888ce7/download/d085e2f8d0b54d4590b1e7d1f35594c1pediacitiesnycneighborhoods.geojson))

**Note:** This project will only consider the average price of 2-bedrooms condo, which is a common type of real estate among normal families.

**1. Data collection process:**

* The average price will be scrapped from the CityRealty website.
* For each neighborhood, call Geocoder Python 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 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 average 2-bedroom condo price of that neighborhood.

**2. Using data to solve the question:**

* First, correlation between price and surrounding venues will be checked.
* Second, if correlated, machine learning techniques (PCA, Regression, PCR) will be used to analyze the data. The output will be a list of venues types that effect the most on price, along with their weight on the result.