**Data Science Capstone Project Report**

**Introduction/Business Problem**

Project is to explore business opportunities in the luxury segment in the US. His aim is to find places with good of the market preferably in the metropolitan areas. He wishes to target the very high-income section of the society. Need to help find good spots to start the new business and also provide some constructive market advice about the neighborhood and potential scope.

This case can serve as a lookup model for a target audience

Data

The preferable regions to start the work would be in the metropolitan area. So the more specific objectives would be to collect the data of the different metropolitan areas in the US, analyzing the different parameters. Because we need to identify the very-high income class, we choose the target market as the people whose income is in the top 1% of the metropolitan area. We would need to identify the mean income of these people and size of this 1%. Use dataset as below:

* Ratio of top 1% income to bottom 99% income for all U.S. metropolitan areas, 2013 (Source: Internal Revenue Service SOI Tax Stats (various years), and Piketty and Saez (2012). Core Based Statistical Areas defined by the U.S. Census Bureau, Population Division; Office of Management and Budget, February 2013 delineations)

For example, a part of data looks something like:

Metropolitan area   Average income of the top 1%       Average income of the bottom 99%       Top-to-bottom ratio   
Jackson, WY-ID                  19,995,834                                                   93,891                                                   213.0   
Bridgeport-Stamford, CT     6,061,230                                                    82,222                                                      73.7

Here we can get a clue of what do elite members of the particular metropolitan area look like. We can also look at the population data separately to check if this 1% is significantly big.   
Hence we can choose an area with significantly rich people.

Finally, the location of this area is identified and analyzed in Foursquare data.

We would use the Foursquare location data to explore the nature of the neighborhood market in any one of the top possible places. And pop some examples that can be used as role models or case study to start one’s own business

For example, which are the most popular fashion brands in this area and where specifically are they located. This can help us make strategic decisions.

Methodology

**Extraction**

To execute the analysis, we first need to extract the data of the mean income from the data source into a pandas dataframe.

**Refining**

Following this we filter out unnecessary columns like top-bottom ratio ranks and the missing data.

**Processing**

Now we sort this data to identify the richest spots. We select a list of the top 30 metropolitan areas. Jackson, Stamford and Marco Island top the list without knowing.

We refer to metropolitan dataset. The source of the data was [*https://www.currentresults.com/Weather-Extremes/US/largest-cities-list.php*](https://www.currentresults.com/Weather-Extremes/US/largest-cities-list.php)

Now we convert this into a dataframe, then we look at the metropolitan areas in the US with a population more than 3 million. This filter to the dataframe we get a bunch of 17 most populous cities in the US.

Next task is to identify which of these cities are also in the list with a rich top 1% of the population. Once that is done, we select the best choice from both dataframes by putting certain level of thresholds as filters. We get the metropolitan area of our choice.

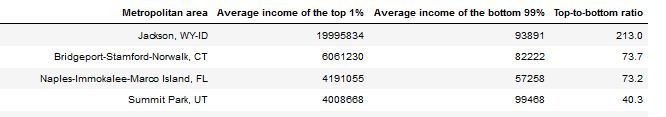
Neighborhood Analysis

We know what our target area, we look at the popular localities/neighborhood on the net and a quick google search suggests what the popular areas are. We choose one of them as the search location and look for places within a preferred radius, let’s say 1000m. The limit was set to 400.

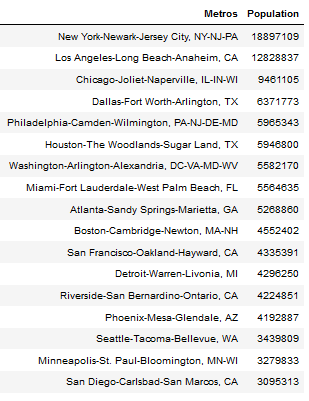
These actions on the Foursquare dataset will tells us the businesses and specific locations.

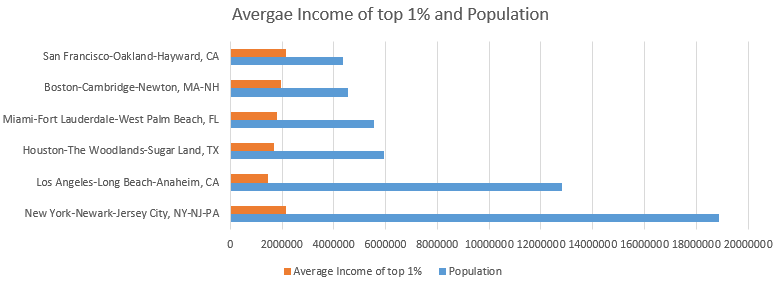
**Results**

The results after extraction and refining looks as below:

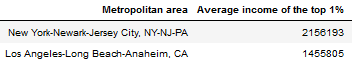


We have a dataframe cleaned, aligned and sorted in descending order with the top 30 metropolitan areas of mean income of the top 1%.

In the processing stages, we extract the metropolitan area wise population data & select the areas with population>3 million. 



Now we have the information about the top spots. It seems that New York and Los Angeles are clearly the best places with population>10 million. We compare both population and mean income tables for New York and Los Angeles.



New York being slightly greater than LA in terms of population. Also New York is slightly better than LA in terms of mean income of the top 1% as well.

Therefore, we choose New York. A quick google search of the posh shopping areas in New York reveals that Upper East Area is one of the most popular.

So we make further Foursquare queries about the sort of categories of shops and location streets and avenues where these stores are. The lists are as follows.

Kind of businesses:

* 'Accessories Store',
* 'Art Gallery',
* "Women's Store",
* 'Optical Shop',
* 'Design Studio',
* 'Cosmetics Shop',
* 'Nail Salon',
* 'Furniture / Home Store',
* 'General Entertainment',
* "Men's Store",
* 'Nightlife Spot',
* 'Jewelry Store',
* 'Office',
* 'Event Space',
* 'Miscellaneous Shop',
* 'Boutique',
* "Doctor's Office",
* 'Co-working Space',
* 'Clothing Store',
* 'Business Service',
* 'Convention Center',
* 'Indie Movie Theater',
* 'General College & University',
* 'Exhibit',
* 'Pet Service'

Popular locations for businesses within:

* '3rd Ave',
* 'Between 5th & 6th Aves',
* 'Between 6th & 7th Aves.',
* 'Madison Avenue',
* 'Broadway & Amsterdam',
* '45th Street'
* '47th Street',
* '48th St',
* 'Between E 49th & E 50th St.',
* '51st Street',
* 'Between E 56th & E 57th St.',
* '58th Street',
* 'E 60th St',
* 'W 63rd St',
* 'W 74th St',
* 'Between 78th & 79',
* '82nd Street',
* 'Between 87th & 88th'