**Coursera Data Science Capstone Project Report**

* *Aditya Mahajan*

**Introduction/Business Problem**

A person is trying to explore business opportunities in the luxury segment in the United States. His/her goal is to find places with a decent chunk of the market that is interested in such high-end luxury goods or services, preferably in the metropolitan areas. More specifically he/she wishes to target the very high-income section of the society. We need to help find good spots to start the new business and also provide some constructive market advice about the neighborhood and potential scope. The motivation is let the user make aware decisions based on smart analytics. *(Note: The cost of setting up business is not a concern in this case - We only have to identify high potential space)*

This case can serve as a lookup model for a target audience that is exploring market opportunities in a region and trying to find potential areas for opening stores for a particular commodity by identifying areas abundant in people who are target buyers for the product. Such neighborhood analysis can help identify opportunities of possible new businesses that can be started in an area based on the income distribution and purchasing power of the residents. On a broader outlook, it can act as a tool for combining huge scale of data to make smarter decisions while choosing best spot to setup businesses.

**Data**

The preferable regions to start the new work would be in the metropolitan areas. So the more specific objectives would be to collect the data of the different metropolitan areas in the US, analyzing the parameters like income. Since 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%. We would be using the dataset mentioned 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)

It mentions the average income of the top 1% of the population in each metropolitan area in the US. Based on that we can identify the places with possibilities to start a new venture. We might additionally look at other parameters as well depending on the results.

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 while identifying spaces with high potential. People might prefer spaces with many similar shops or some might prefer to set up their shops slightly away from their rivals but close to other shops that sell complimenting products. A store selling jewelry may like to be close to a store that sells high end designer clothes. We can help identify such depending on the preference and choices.

**Methodology**

**Extraction**

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

**Refining**

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

**Processing**

Once we have a clean data, 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 but we don’t’ know if these places have a reasonable market size.

We therefore refer to metropolitan dataset on the net and use input it as a list. 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)

First we convert this into a dataframe, then we look at the metropolitan areas in the US with a population of more than 3 million. Applying 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. Doing this we get the metropolitan area of our choice.

**Neighborhood Analysis**

Now that 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.

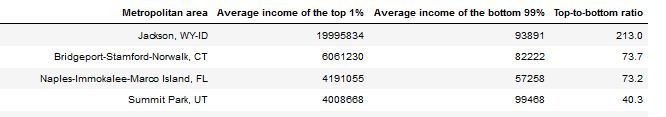
We can try to look up for as many luxury segments as we want like luxury cars etc. However, for this project lets choose the fashion sector. We look up the Foursquare database to identify the relevant spots in and around the selected subarea.

Once we get the data we refine and polish it to remove missing value entries and also to make the data more readable in terms of categories and locations, we get are able to identify the sort of fashion stores, streets and avenues with popular fashion stores

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

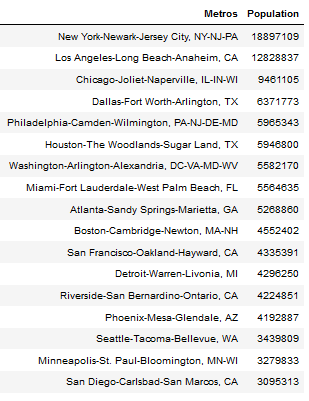
**Results**

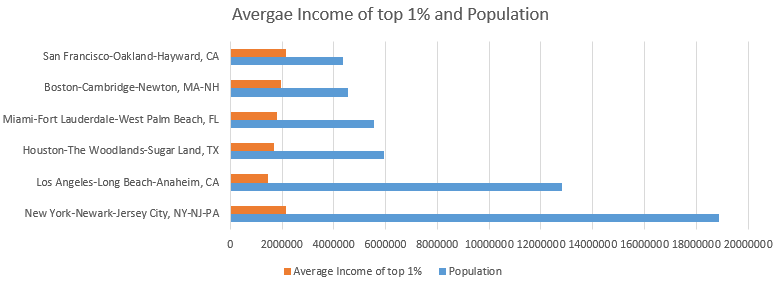
The results after extraction and refining looks as follows:



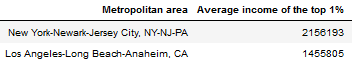
We have a dataframe cleaned, aligned and sorted in descending order with the top 30 metropolitan areas of mean income of the top 1%.

In the processing stages, we extract the metropolitan area wise population data & select the areas with population>3 million. We get a bunch of 17 most populous cities in the US, which are as follows.





Now that 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'

**Discussion**

Hence our client now has a pretty refined set of choices and options which came out as a result of our thorough data analysis. The client can now decide what businesses have scope of opening up in this region and where is are the best places. Of course, a more detailed study of customer opinions and market demand and history would be vital from the perspective of long term profitability and other market choices. This study was aimed at providing a way to identify the hubs or hotspots for a particular industry or market. In our case it was the luxury fashion segment. We first scraped data, then played with it to filter out some relevant operations. Following this, we used the tools like Pandas and Foursquare to get a high definition picture for our choices and decisions. It clearly highlights the vitality and benefit of using these tools for data analysis. We are able to derive meaningful insight from a crude dataset and help people to make smarter choices with strong evidences rather than the popular hearsay.

**Conclusion**

Our method thus provides our client access to set of choices that meet his or her needs and can also hint him or her to choose the sort of businesses that should be set up according to the place. It’s not the final decision making process but it’s the process that would support the final decision making process and make informed choices. As per our analysis the client should start the business in the New York –New Jersey metropolitan area. The set of locations and businesses prevailing there were found. In fact, some of results are very similar to the top net searches like suggestions of places like 5th Avenue, Broadway etc. However, some of the suggestions in businesses were really interesting and might not be very obvious when searching the net like Universities about Fashion, Indie Movie Theatres etc. These could help our client make the choices in a more informed manner and start a new business with much more ease.