- **INTRODUCTION**

  This project will provide an analysis on possible strategies about HORECA business investment location selection in Tokyo metropolis. To be more precise the analysis should provide some better understanding on possible choices for investment in new restaurant location/opening.
  Each end every investment – whether it's about buying stocks, starting new business, company assets or any other – if it's about spending money it should be ideally the result of a strategic and broad analysis. The better understanding of the planned investment an investor will have the higher are chances that the final result will meet and fulfill the expectations.

- **STRATEGY**

  Once the decision about an investment has been made it should be followed by the analysis. One should keep in mind that in most cases there is no guarantee that the result will be success – no matter what will be the definition of success and how we would like to measure the success, yet applying some analytical tools and following a structured strategy of decision process we can increase these chances for the possible future success. In most cases decision makers and their teams are using one of the standard strategies (approaches) – "top-down" and "bottom-up". While looking for some general rules or trends we usually use bottom-up approach (form details to general conclusion), then searching for a specific solution to a given problem will probably involve top-down approach.
  For the use of this project I decided to apply kind of strategy that for my own purpose I will call "walking-the-rainbow". This fancy name somehow is giving at glimpse the sense how the work should evolve. So let's start by scratching the surface.

- **PROBLEM AND HOW DO I INTEND TO DEAL WITH**

  Depending on the investment's nature there can be many different types of analysis applied just to name some basic problems like – legal aspects, profitability, feasibility and so on. Some of them do not require numerical/computational analysis (f.i. legal aspects) but most of decisions we would preferably like to take based on some kind of "hard and solid" proof or evidence. In most cases this implies the use of computational methods. If we imagine just a very simple case decision tree with couple of decision nodes levels we will see that the number of possible solutions can grow dramatically and it would be very difficult (or maybe impossible) to analyze each possible scenario without the use of some sort of data science. Even though not all input data for the analysis must be numerical (some of them can be categories f.i.) still by applying ML methods we can try discover some general rules, trends or best possible scenarios for a given task.

  Considering our given case – investment in HORECA (HOtel REstaurant CAfeteria) let's get some better feeling and let's decompose the problem into some basic steps:

- scale of business (restaurant) – for example just three categories – small, medium, big
- type of restaurant – fast food, luxury, family, vegetarian, casual,...
- basic cuisine – Italian, Greek, French, American, Japanese, Chinese, Indian...
- sources of financing and level of leveraging (if the case)
- financial indicators like expected IRR, NPV, DCF, ROI, PB
- exit scenario
- sensitivity analysis – trying to identify which factors can possibly affect the financial result of our investment
- last but not least – location selection

The above very short list of problems to be considered can be adopted into a decision tree and will give us at least seven levels of decision nodes. Above picture with very limited branches at each decision node – numbers on each level represent possible branches for each node on this level – is showing the scale of the problem. Using the numbers form the list above we will reach… 24,300,000 possible solutions! For the need of this project I'll just focus on the problem of location selection applying as mentioned above a "walking-the-rainbow" strategy. Why do I use this name? I assume that we start from a simple task – select quasi-optimal location for the restaurant in Tokyo. This will be one end of the "bow". Than we will climb the rainbow collecting available and hopefully essential data, creating broad (global) picture of possible location and finally I will walk down the rainbow trying to pick couple of most perspective location. This kind of analysis is important for every investment location in the world, while in this specific place called Tokyo we must keep in mind that Tokyo is ranked to be the most populous metropolitan area in the world (Greater Tokyo Area) with over 38 million people. This is also the world's largest urban agglomeration economy which is hosting (as of 2011) 51 of the Fortune Global 500 companies (the highest ratio for any city of the world). Tokyo ranks first in the Global Economic Power Index and third in the Global Cities Index.
Given just this few basic indicators the stated problem of location seams to be very crucial for the planned investment to be successful. Being so I will try to find out which are the most popular venues for the "lunch and/or evening out of home" and which of them tend to be more diverse with regard to the type of cuisine/restaurant/coffee shop while (at the same time) if there are any other neighborhoods which seem to be more traditional regarding offered types of cuisine (since we are discussing Tokyo it would mean locations where majority of food offering is Asian type food). Afterwards will try to elaborate where could be a good location and what could be the best choice for this investment.

- **DATA**

One of the data sources will be FOURSQARE service provider. I will use basic infos on venues (categories) related to places predefined on the basis of ZIP numbers.
The second source of information (to get ZIP numbers for given locations in Tokyo) will be a web service of AggData (www.aggdata.com/free/japan-postal-codes) which once I downloaded the required zip-data it was reduced to zip codes of selected part of Greater Tokyo Area.

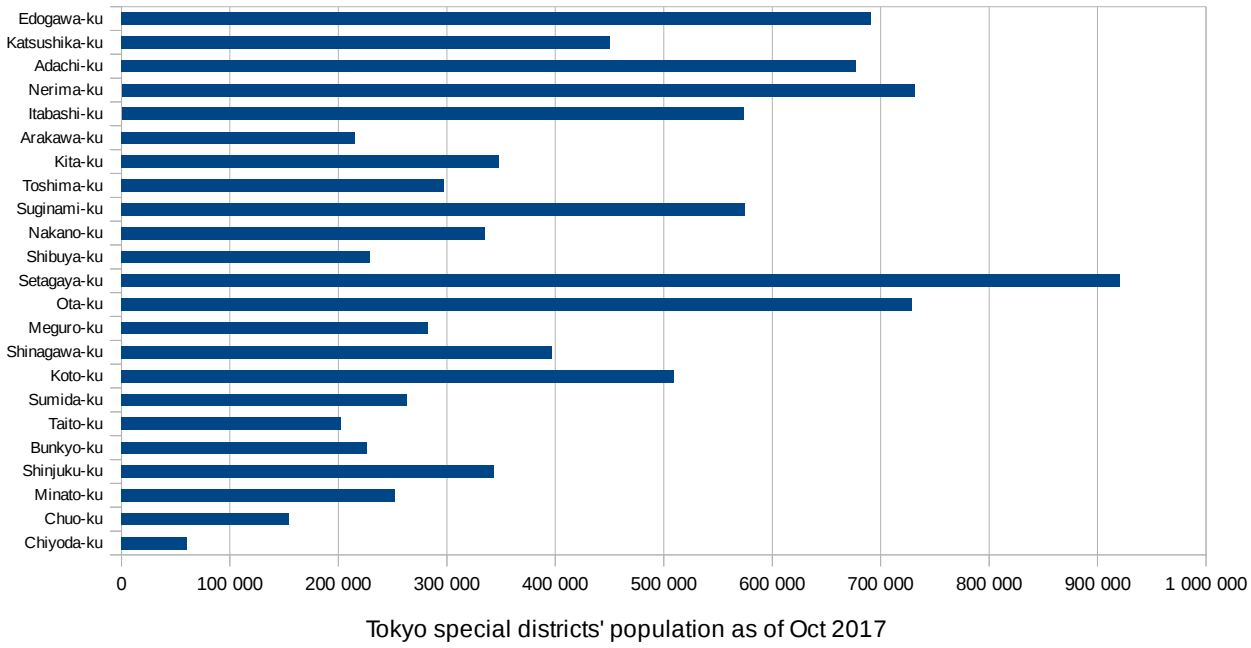
The third source of data (statistical data) will be official web service of Tokyo city and it's dedicated statistics part www.toukei.metro.tokyo.jp

- **METHODOLOGY**

One of the important tasks was to understand the administrative structure of this very unique metropolis and to gather and prepare some statistical background for the further work. This part of the work was based on data provided by the official Tokyo source. The very first topic was to distinguish the Greater Tokyo Area from the Tokyo urban area. This is the consequence of understanding that the actual Tokyo metropolis was formed by merger of the Tokyo prefecture and the original Tokyo city. As the result today Tokyo is a combination of 62 administrative districts which can be somehow referred as Boroughs in other big metropolis. Yet the original, so called "special" (tokubetsu-ku) wards of the city of Tokyo are limited to 23. Still the number of population of this part of the Tokyo metropolis is around 10 million people which will be the focus of the project and the challenge for some analysis. Having this decision done I have collected demographic and statistics data from the official Tokyo city website as described in the data section. Bellow are presented some numerical and graphical forms of this data.

- **POPULATION BY DISTRICT as of 2017**

District	Oct. 1, 2017 (Estimate)
千代田区 Chiyoda-*ku*	60 934
中央区 Chuo-*ku*	154 728
港区 Minato-*ku*	252 786
新宿区 Shinjuku-*ku*	343 067
文京区 Bunkyo-*ku*	226 419
台東区 Taito-*ku*	202 462
墨田区 Sumida-*ku*	263 484
江東区 Koto-*ku*	509 438
品川区 Shinagawa-*ku*	396 993
目黒区 Meguro-*ku*	282 785
大田区 Ota-*ku*	728 349
世田谷区 Setagaya-*ku*	921 120
渋谷区 Shibuya-*ku*	229 519
中野区 Nakano-*ku*	335 377
杉並区 Suginami-*ku*	575 326
豊島区 Toshima-*ku*	297 763
北区 Kita-*ku*	348 425
荒川区 Arakawa-*ku*	215 868
板橋区 Itabashi-*ku*	573 669
練馬区 Nerima-*ku*	731 082
足立区 Adachi-*ku*	676 761
葛飾区 Katsushika-*ku*	450 014
江戸川区 Edogawa-*ku*	691 121



Tokyo special districts' population as of Oct 2017

Next step was to identify if there are any “preferable” locations for restaurants in analyzed wards and to discover, if possible, what were the reasons for any preferences. The results of this part of the project are presented in the next paragraph.

- **RESULTS OF APPLIED DATA ANALYSIS**

Based on zip codes of the original city of Tokyo I retrieved from the Foursquare data on venues and their respective locations. Originally I had 64,586 zip codes of Japan. The first step was to reduce this set to the Tokyo metropolis which resulted in the set of 1735 of zip codes with above mentioned 62 unique counties which I decided to limit just to 23 special wards of the original (historical) Tokyo city. This brought me to the set of 1032 unique zip codes. By using Folium and geodata from this set I created a first picture of Tokyo with Counties superimposed on top. This is very stunning and impressive first view. Next I retrieved from the Foursquare data of venues. The result was 48,004 venues with names, categories and respective geo coordinates. Afterwards I applied similar clustering ML like used for NY and Toronto analysis, gathering most popular types of venues for each given neighborhood. This revealed some information if the neighborhood is a popular leisure destination and if it has kind of international character (various types of restaurants/cuisines – pizza, French, Burger etc) or if it tends to be more traditional focused, say limited to Asian types of food/restaurants. Here are basic steps of the applied methodology:

- 1. Prepare dataframe (DF) with zip codes for selected counties of Tokyo metropolis (original 23 “Tokubetsu” counties (“ku”) → result DF “Zip_tokyo_tokubetsu_ku”. This DF contains 1032 unique zip codes for areas in preselected 23 counties.

- 2. Get from Foursquare information on Venues corresponding to the list of zip codes → result DF “Tokyo_venues”. This DF contains basic information on 48004 unique venues (yet this value is changing almost every day as new venues appear in the Foursquare database).
- 3. Rearrange venues in the way which can bring information which county could be good choice for the HORECA investment. This means group venues by at least “Place” which corresponds to the unique zip code → result DF “Tokyo_venues_counted”. This DF contains information on 886 different (unique) places (locations – based on zip codes) which correspond to the original 48004 unique venues (stores, restaurants etc.) collected from the Foursquare service.
- 4. Do clustering on special DF – 886 “places” and 444 “categories” of venues → “Tokyo_grouped” DF with k-means clustering → result DF “Tokyo_grouped_clustering”
- 5. Finally select the cluster with the highest number of places (locations) that obviously correspond to the highest density and variety of “venues” - in our case this is cluster number 6 which “collected” 303 various places (zip codes).
- 6. Having that done I’ve sorted places of the cluster 6 in descending order by the number of “venues” bound with each “place” (zip code location)

- **DISCUSSION OF RESULTS**

  This analysis finished with clustering revealed couple of interesting information – first there is a vast amount of Tokyo venues in the Foursquare database and out of these 48T venues majority points of interests (for tourists) are exactly restaurants, bars and different kind of HORECA locals. Next information we can draw off the analysis is that there exist a huge variety of HORECA locals in Tokyo Metropolis – Ramen restaurants, traditional Japanese restaurants of different “specializations” through Chinese, Italian, French up to Indian and Thai restaurants. Finally clustering of these venues “concluded” that the surroundings of parks - in Japanese word “kouen” means park (Uenokouen, Komazawakouen, Heiwanomorikouen, Kitanomarukouen…) are the most popular locations for plenty of venues. And Indeed Ueno Park located next to Ueno JR station is one of the very top and famous for cherry blossom viewing (so called Hanami) with millions of visitors every year – so no surprise that it is the very first venue location to be considered for HORECA ivenstment.

- **CONCLUSION**

  As the final conclusion for this very specific investment I would say that such extraordinary cities like Tokyo, such unique culture and business approach and a specific mentality of citizens aside of the technical approach and computer data analysis demands from Data Scientist (or even Data Science Team) additional background knowhow (sometimes even language knowledge). Of course this business case which I presented even though fictive could be or in fact it is a daily investment problem and should be followed by additional financial analysis – fi. locations I were able to preselect should be investigated on ground prices vs planed budget for investment, availability of parcels and if so than the size of available locals or parcels should be also scope of interest to not mention strict financial

aspects like NPV, IRR, PBT and many more yet this does not directly involve Data Science or any form of AI, ML etc.