

When Data Science is applied in Business Decision

Data Science is a recent umbrella term that includes machine learning (ML) and statistics, certain aspects of computer science including algorithms, data storage, and web application development.

This is my IBM Data Science Capstone Project for the last milestone of IBM program in acquiring **Professional Certification in Data Science**. In this project/assignment we are asked to build a showcase where we will clearly define a problem and discuss the data that we will be using to solve the problem. The solution should leverage the use of **Foursquare location data** to explore, solve or execute the business case. Python codes for Business Case data Analysis & visualization are given in the attached Github link.

1. Introduction

Guaranteed business quality improvement, cost reduction, agility & flexibility to scale up/down the business, and managing customer experience are all key reasons for business owners to outsource some activities (managed service) to focus on their core business. Managed services can be, but not limited to:

- Business quality improvement (KPI – KQI – CEI)
- Preventative Maintenance programs
- Equipment installation
- 24 Hour Emergency Service & Call-Out Support
- Spare Part Management
- Work Force Management (WFM)

Nevertheless, within the restaurant business industry, preventive maintenance of equipment and systems are overlooked. Restaurateurs often choose to maintain their equipment on a reactive maintenance or breakdown model. A good approach to raise some concerns for those who adopt this strategy is to use capital budgeting technique and calculate the ROI, NPV, IRR, PBT...etc. over short and long term. From this perspective, we can evaluate reactive vs. proactive maintenance approach.

1.1. Business Problem

One of my clients has been in the managed services for **Restaurant Preventive Maintenance business** for almost 2 years. His first outsourcing service branch in “Cairo - El Agouza” neighborhood has shown a promising level of financial indicators in terms of ROI and NPV. From this perspective, he believes that establishing a second office branch for restaurant managed service is the way forward to grow the business. My client’s request was to support him in developing the new business plan document and to do the business case model (VBM), which I did. It is worth noting that while working on the business plan for the new branch I had to do **data analysis, visualization, and machine learning (ML) with Python** to answer some of the following questions:

- **First:** Which neighborhood to pick to open the new office for restaurant managed service? The office main goal is to serve venues (restaurants & cafés) in this neighborhood mainly. Selection should consider a number of criteria; the cost of property investment in the selected neighborhood must result in positive NPV, venues are surrounded with particular household income segment, and the neighborhood should be within a maximum geo-location distance of 12km from his first office in “El Agouza” where travelling time plays an important role between his first and second office.
- **Second:** Due to office resources limitation, how to screen or filter the potential venues list (restaurants, Café...etc.) in the neighborhood to identify target venues (customers)? In other words, what are the criteria to be used for screening?
- **Third:** For an efficient Inventory Management for Spare Part Management Service (SPMS) portfolio, how much set of items to order monthly (QTY) based on historical data of “Supply and Demand” pattern reported from his first office in “El Agouza”?
- **Fourth:** For an efficient Inventory Management, how to find out the actual number of each item to invest when there are certain constraints between two items? For example, for every 2 units of the Item (A) consumed, 1 unit of the Item (B) consumed, the total volume (A + B) must not exceed 102 units according to his historical data.

1.2.Target Audience

- An entrepreneur who wants to start his business in managed services. Business location is an element of 4Ps marketing mix, therefore which neighborhood is most likely to have restaurants (Fine Dining, Casual Dining, Family Style, Fast Casual) generating good profit and congested with people should be investigated.
- An Inventory Capacity Planner. Similar to Economic Order Quantity (EOQ) he wants to estimate the order quantity at one time with degree of certainty. With limited data for item order rate, what kind of probability distribution function (pdf) to use with certainty?

2. Input Data

2.1.Household Expenditure at Restaurants

There are many factors that will contribute to restaurant's success including food quality, service quality, marketing, competition, and location. Business location plays an important role in terms of the number of restaurants already exists and category of surrounding households average income.

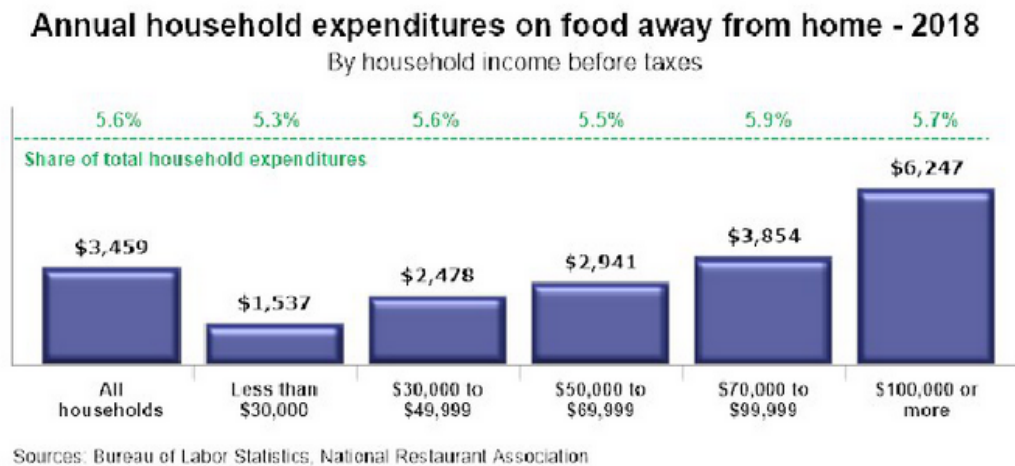
Expenditure Survey data from the Bureau of Labor Statistics in US shows that spending at restaurants and other foodservice outlets on meals, snacks and nonalcoholic beverages represented **5.6 percent** of the average household's total expenditures in 2018. The survey also indicated that spending at restaurants rises along with household owner income. This means we want to look for locations where the median ***household income is close to or above certain level in order to address restaurants with successful business (ROI).***

```
import matplotlib.pyplot as plt
from PIL import Image # converting images into arrays

# download image
!wget --quiet https://restaurant.org/Restaurant/media/Research/Econ%20Notebook/Consumer-Expenditures-Income.jpg

# save mask to household_exp
household_exp = np.array(Image.open('Consumer-Expenditures-Income.jpg'))

fig = plt.figure()
fig.set_figwidth(14) # set width
fig.set_figheight(18) # set height
plt.imshow(household_exp)
plt.axis('off')
plt.show()
```



<https://restaurant.org/Articles/News/Households-in-the-West-spent-the-most>

2.2.Average Residential Property Prices

From “Global Property Guide” in Egypt, I will investigate the **Average Residential Property Price** per square meter in every neighborhood to achieve two goals:

- To comply with project accept/reject criteria for the developed business case by identifying the suitable property price/Sqm. For an office space area of 60 to 80 square meters the investment cost in a condominium **should not exceed (EGP 8k/Sqm)** to get an approved capital budgeting measures (ROI, IRR, NPV, PBT...etc.)
- To address a particular household owner segment as there is a positive correlation between property value and annual household owner income. Moreover, a study showed that household spending at restaurants rises along with household owner income. Therefore to cope with the household spending considered in the business case, the Residential Property Average Price/Sqm **should be greater than (EGP 6k/Sqm)**.

To scrap the “Residential Property Prices” site I will use Request library and BeautifulSoup to fetch and parse the data. Then I will sort the parsed table content with price per square meter.

<https://www.globalpropertyguide.com/Middle-East/Egypt/Price-History>

2.3. Geographical Coordinates of Cairo Neighborhoods

One more important aspect for the new neighborhood selection is how far it is from the first office branch in "El Agouza"? Because the Service Level Agreement (SLA) has strict resolution time for Emergency & Call-out support it is important that the new office is relatively close to first office in case of support is needed in either branch. This will help to avoid penalty or contract breach. The geo-location distance from the first office should not be greater than 12km.

Working with "geopy.geocoder to import Nominatim" will provide the geo-location of neighborhoods (Latitude & Longitude). Then using "geopy.distance" will help to calculate the geodesic distance from the first office in "El Agouza" to each neighborhood in the list. The geodesic distance is not the road/route distance but it is the shortest distance on the surface of an ellipsoidal of the earth.

Scraping the web site of "neighborhood property price" will give us Cairo neighborhood names to get its geo-location

<https://www.globalpropertyguide.com/Middle-East/Egypt/Price-History>

2.4. Foursquare API for Venues Exploration

Once the target neighborhood for the second office meets the three selection criteria (Capital budget, the particular household owner segment, location distance), foursquare API will get the top 100 venues that are within a radius of 1,500 meters. Data is taken from JSON file under ['response']['groups'][0]['items'] to be structured into a pandas data frame.

```
url =  
'https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&ll={},{}&radius={}&limit={}'.format(
```

2.5. Municipal Divisions of Cairo Governorate with Population

Just to have the feeling of how condensed the selected neighborhood, a list of all the municipal divisions in Cairo city with an estimated population for each division from wikipedia. Scraping web table using "pandas.read_html" function will manifest all Cairo neighborhoods with population in Data-Frame.

https://en.wikipedia.org/wiki/Cairo_Governorate

2.6. Inventory Historical Data

This is last year monthly warehouse report "Supply & Demand" items. The report listed quantity of combined items only with no show to each item quantity separately (see the table below). This is due to the nature of equipment where some spare part items are not replaced alone or rarely are replaced separately.

	(2a + b)	(a + b)
Month		
1	70	60
2	85	80
3	85	80
4	90	80
5	85	65
6	90	82
7	85	78
8	90	85
9	95	87
10	85	77
11	85	80
12	120	90

When applying (solving with) linear programming technique (LP) on the historical data you can get the ***minimum quantity for each item*** to minimize inventory cost (objective) while meeting the demand of spare part items (constraints) according to reported data (explained in details in later sections). Since the spare part items' quantity varied over the last 12 months with no clear sign which distribution it followed, Triangular probability distribution will be used to set the quantity for the above constraints with likelihood of occurrence (CDF) of not less than 80%.

Using python, the following is applied "from scipy.stats import triang" and "import pulp"