# Location selection based on data science techniques

Capstone assignment under the umbrella of the

Coursera / IBM Data Science Professional training

Author:

Roger Moll

June 19th, 2020

# Introduction

Cities are very diverse and are the financial capitals of their respective countries. To be close to the markets and customers, companies are put into challenging situation to make the best choice. Where shall the future regional headquarters be to fit best into a companies purpose. For this decision, multiple factors such as taxation, labour and skill abundancy, political stability among others are essential decision variables.

For this assignment it is assumed that we advice a company in their decision process of choosing a location for their Europe headquarter. The company has pre-selected three cities in which a local office already exists:

* [London (GBP)](https://en.wikipedia.org/wiki/London)
* [Zurich (CHE)](https://en.wikipedia.org/wiki/Z%C3%BCrich)
* [Barcelona (SPA)](https://en.wikipedia.org/wiki/Barcelona)

Each of these cities are attractive places to live and work. All are very close to international airports and are appreciated by Expats. The organisation wants to settle his new headquarter based on the following decision matrix:

A screenshot of a cell phone

Description automatically generated

Figure 1: Decision Matrix show importance of decision variables

A special focus is given on "Gen-Y" suitability as this will become both the predominant customer base as well as the main source of future employees and leaders.

GenY or aka "Millennials" is defined as the demographic cohort following with the early 1980s as starting birth years and the mid-1990s to early 2000s as ending birth years [Wikipedia](https://en.wikipedia.org/wiki/Millennials), have the following characteristics:

* Millennials are tech-savvy as grew up with technology, and they rely on it.
* Millennials are family-centric and are willing to trade high pay for fewer billable hours, flexible schedules, and a better work/life balance.
* Millennials are confident, ambitious, achievement-oriented but have high expectations towards their employers and aren't afraid to question authority. Generation Y wants meaningful work and a solid learning curve.
* Millennials are team-oriented. They want to be included and involved.
* Generation Y craves attention, feedback and guidance.
* Generation Y is prone to "job-hopping" as they're always looking for something new and better.

An interesting summary is provided by [Goldman Sachs](https://www.goldmansachs.com/insights/archive/millennials/). However, any new location should be "cool" with this Generation's life style and preferences' as to offer a natural cause to remain and limited need for location moves.

# Data

## Data sources

The sources of data acquired for this assignment are well known and regarded NGO institutions as well as data collection platforms e.g. Foursquare. The following data sets were collected

|  |  |  |  |
| --- | --- | --- | --- |
| Name of organization | Data access /  Report | Data attributes | File format |
| Foursquare | API | Id, name, contact, location, categories, verified, stats, url, hours, popular, menu, price, rating, hereNow, storeId, description | API (json) |
| OECD | [Unemployment](https://www.oecd-ilibrary.org/economics/data/labour/labour-force-statistics_data-00046-en) | Annual, harmonised unemployment rate in %, over 10 years | API (csv) |
| OECD | U[nit labour cost](https://stats.oecd.org/Index.aspx?DataSetCode=ULC_EEQ) | Annual, Unit labour costs and labour productivity (employment based), Total economy in %, over 10years | API (csv) |
| OECD | [Corporate Tax](https://stats.oecd.org/Index.aspx?DataSetCode=TABLE_II1) | Statutory corporate income tax rate, Total economy in %, over 10years | API (csv) |
| OECD | [Labour wage](https://stats.oecd.org/Index.aspx?DataSetCode=EAR_MEI) | Average annual wages, Total economy in EUR, over 10years | API (csv) |
| Transparency International | [CPI Data Set](https://images.transparencycdn.org/images/2019_CPI_FULLDATA.zip) | CPI rank, CPI score, CPI std. error | XLS |
| Opendata.swiss | [Swiss city index](https://opendata.swiss/en/dataset/amtliches-ortschaftenverzeichnis-mit-postleitzahl-und-perimeter) | Administration, Construction and housing, Crime, criminal justice, Culture, media, information society, sport, National economy, Education and science, Energy, Finances, Geography, Mobility and Transport, Public order and security, Politics, Population, Prices, Social security, Health, Territory and environment, Tourism, Work and income | API (json) |
| Stadt Zurich | [District information](https://data.stadt-zuerich.ch/dataset?res_format=json&tags=polygondaten) | Geospatial data as polygon lat, long coordinates | Json |
| Statista | [Average price of residential property](https://www.statista.com/statistics/722905/average-residential-square-meter-prices-in-eu-28-per-country/) | Average prices of 120 square meter apartments located in the most important cities of 38 European countries | Web scraping |

# Data cleaning

The data downloaded or scraped from multiple sources were put into single tables for further processing. The data sets need to be:

* put into consistent data types
* Checked for missing data, i.e missing values for single years were added by building the average over the existing data. Series which were missing entire rows / set of data were deleted.
* Despite a standard download and parsing of the data sets (UTF-8), special characters needed to be replaced and locations be standardized
* Most data set needed to be transposed for the graphical rendering and to have a common key (e.g. years)

At a later stage the multiple data set were conjoined into one large table to allow a decision tree to be deployed.

A screenshot of a cell phone

Description automatically generated

Figure 2: Subset of the gathered country information

The data statistic revealed complete, yet small data sets of 30 tuples. I’ve decided against further resampling to enlarge the data set for time reasons as the collection and harmonization of the data from the various sources was time intensive.

A screenshot of a cell phone

Description automatically generated

Figure 3: Statistics of the collective data set

# Methodology

I’ve decided to approach the issue statement – which is the best location to set up a new Europe head quarter – in two steps.

1. Select a country among the three candidates (SPA, CHE, GBR) based on macro economical data as shown above and run a decision tree analysis on the subject to define the key selection points i.e. the high entropy gains.
2. Based on the preselection, and with the data of mainly Foursquare, the different boroughs were analysed for the suit to the companies need and GenY compatibility. For this a clustering and prediction model was deployed on the data

Step 1:

Looking at the macro economical data e.g. labour cost, unemployment rate, corporate taxes and productivity (ULC), provides a solid view on how favorable the economic conditions in a country are. This allows a robust preselection.

Caution is given and the observer needs to be aware that these are macro economical views and the actual situation in a city’s microenvironment might be different. However, since all of the data are harmonized across the countries, we measure comparable data and trends.

Observations:

|  |  |
| --- | --- |
| A close up of a map  Description automatically generated | A close up of a map  Description automatically generated |
| A screenshot of a map  Description automatically generated | A screenshot of a map  Description automatically generated |

Interpretation of the data

* Development of productivity (ULCs):  
  All three candidate countries show an upward trend in ULCs which is in the end a deterioration of the productivity.   
  Result: Switzerland, show preferable conditions.
* Development of the unemployment rate:  
  GBR and CHE show unemployment rate below the OECD average. This allows two assumptions: a) both countries will see immigration of labour force as additional capacity is on demand on the countries labour market and b) based on the labour cost, we can assume that both countries, based on their GDP per head, can afford comparatively high skilled labour which is what we’re looking for.  
  Result: GBR and CHE show preferable conditions.
* Development of the Corporate tax schemes:  
  For the economic benefit of a company, it is desired to have fair and low corporate taxes in percent of revenue as this has a direct impact on your net result. Yet as companies with a high degree of corporate social responsibility, it is a mandate to pay taxes in the community located, i.e. tax heavens are not option.  
  Result: Whilst Switzerland shows a stable and hence reliable taxing scheme, GBR has recently lowered the tax burden significantly. It is yet to see how this evolves post Brexit.
* Development of labour salary:  
  All three countries show a stable development of the average salary. Two elements are to observed: a) Switzerland is significantly more expensive and b) Spain sees a decrease in salary levels.  
  Result: Spain’s workforce is currently earning the least which will increase the pressure on skills drainage towards countries with higher salaries.

A final word to the Governance performance index (GPI) provided by “The World Bank” which measures the rule of law for a country. I’ve decided to use BoxPlot graphs to illustrate the performance of the three countries. The table below holds the ranking and the development of the ranking over the last 10 years. It gives therefore a good interpretation whether political shifts are influencing the economic environment for a company.

A screenshot of a social media post

Description automatically generated

Cost are of essential consideration and looking at the three countries one can easily determine that Spain would be most attractive in terms of property cost.

A close up of a map

Description automatically generated

Figure 4: Property cost per m2 in Europe

Bringing all of this together a data set was created holding all of this information and a decision tree was deployed on the data.

Purpose of the above data review was to identify the data sets providing the highest information entropy gains on the nodes with the least variables to look at. The model indicates that only two variables are to be considered for the location decision among the three countries:

1. Index for “Rule of law”
2. Corruption reliance index

A screenshot of a cell phone

Description automatically generated

Applying this in conjunction with the earlier decision matrix, we’d way the political environment as the most desirable design element, given all others are comparatively the same, and decide for Switzerland / Zurich as the location.

Step 2:

Based on the discussion above, it becomes a question on where, i.e. in which borough, in Zurich the company should seek a location.

Zurich as an international City is subdivided into 11 boroughs as illustrated in the below map. Each of these boroughs have their benefits. Tax wise the boroughs are equal but have differences in quality of live and leisure time offerings.A close up of a map

Description automatically generated

We therefore will now compare the boroughs by building various cluster of venues each borough has and decide which cluster suit best to the GenY needs.

Clusters were built among the following categories of venues as available from Foursquare:

A screenshot of a cell phone

Description automatically generated

As mentioned before, each borough has its specialties and characters. Attached the first five boroughs and their 10 most common venues. A screenshot of a cell phone

Description automatically generated

Building the clusters and comparing it with the GenY expectations, we found cluster number 2 as the one describing the set of desirable environment best. It holds multiple leisure time activities as well as the opportunity for local commuting.A screenshot of a cell phone

Description automatically generated

This allows’ us now to formulate a clear recommendation towards the decision takers.

# Results and recommendation

Based on the available data and the methodology, we’d recommend to locate the new Europe headquarter in Zurich in the borough of Fluntern. This holds multiple benefits such as the proximity to leading Universities, local transport and multiple GenY compatible leisure activities.

# Discussion of the results

This work was highly hypothetical and hence might not hold all data points for a company to consider. It can however lead the process as a template.

The challenge in this work was less based on the statistical validity of the analysis but primarily on the demonstration of the following skills

* Finding, collecting harmonizing data from multiple sources
* Building a data framework which allows the deployment of machine learning techniques, despite the small sample of data
* Application of multiple data visualization techniques, i.e. boxplot, superimposed maps and decision trees
* Handling of at least two machine learning techniques
* Formulation of report demonstrating the skills

# Conclusion:

It took me way longer than expected to conclude the Capstone project. This was mainly caused by facing multiple hurdles to overcome in the data collection process. But it was extremely insightful activity and leaves me with a rich inventory of different data sources.

In case of question and remarks, please do not hesitate to reach out to me under [Linkedin](https://www.linkedin.com/in/roger-moll-2923b0/).