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)
- Zurich (CHE)
- Barcelona (SPA)

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:

Nb.	Attribute	Details	Weight
1.	Labour climate	labour cost, productivity, skills availability	6
2.	Political environment	Effectivness of goverment, policy consistency, corruption	7
3.	Access to Universities	Nb. of Univerities and reputation, ranking, size	5
4.	Quality of life	standard of living, recreation, health	4
5.	Labour and skill aboundancy	Skill levle, unemployment rate, labour unions, wage rate	1
6.	Cost of labour	Productivity, exchange rate,	2
7.	Tax structure	Corp. tax rate, social sec. cost, rgulatory barriers	3

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, 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. 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	Annual, harmonised unemployment rate in %, over 10 years	API (csv)
OECD	Unit labour cost	Annual, Unit labour costs and labour productivity (employment based), Total economy in %, over 10years	API (csv)
OECD	Corporate Tax	Statutory corporate income tax rate, Total economy in %, over 10years	API (csv)
OECD	Labour wage	Average annual wages, Total economy in EUR, over 10years	API (csv)
Transparency International	CPI Data Set	CPI rank, CPI score, CPI std. error	XLS
Opendata.swiss	Swiss city index	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	Geospatial data as polygon lat, long coordinates	Json
Statista	Average price of residential property	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.

	U_Emp	UL_Cost	Wage	Cor_Tax	CPI_Rank	GPI_Acc	GPI_Law	GPI_Qual	GPI_Violo	GPI_Effe	GPI_Corr	Prop_Cost	Country
0	19.875	105.6	41034	30.0	60	85.30806	86.25592	84.21053	33.64929	78.94736	82.38095	5921.47	Spain
1	21.408	103.8	40453	30.0	62	83.09859	85.91549	81.51659	48.34123	81.51659	82.46445	5746.78	Spain
2	24.792	101.2	39302	30.0	65	82.15962	83.09859	78.19905	42.65403	82.46445	83.41232	5577.25	Spain
3	26.117	100.3	39391	30.0	59	78.40376	81.69014	79.14692	46.91943	82.93839	78.19905	5412.72	Spain
4	24.450	100.0	39398	30.0	60	76.84729	80.28846	75.00000	55.23809	84.13461	72.11539	5253.05	Spain

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.

	U_Emp	UL_Cost	Wage	Cor_Tax	CPI_Rank	GPI_Acc	GPI_Law	GPI_Qual	GPI_Violo	GPI_Effe	GPI_Corr	Prop_Cost
count	30.000000	30.000000	30.000000	30.000000	30.00000	30.000000	30.000000	30.000000	30.000000	30.000000	30.000000	30.000000
mean	159.705167	100.703333	49230.566667	23.806667	74.70000	90.810827	90.792382	90.454791	68.360336	90.827791	88.499985	8882.651000
std	816.764407	3.398629	10514.242901	3.773129	11.18851	7.209178	6.653487	8.136289	20.684020	7.056335	10.205483	2757.998435
min	3.767000	95.300000	38761.000000	19.000000	57.00000	76.847290	79.326920	75.000000	33.649290	78.947360	68.269230	4522.610000
25%	4.797750	99.025000	40179.250000	21.100000	62.00000	82.843612	83.802815	81.316648	55.357140	83.413462	82.401825	5619.632500
50%	6.846000	100.000000	44398.500000	21.600000	79.00000	92.409135	93.143875	94.989520	59.619725	91.904580	93.314615	10067.775000
75%	19.045750	101.200000	62932.250000	27.500000	85.00000	98.098935	95.879780	96.153850	94.306027	97.630330	96.409607	10996.052500
max	4484.000000	111.600000	64285.000000	30.000000	86.00000	99.052130	99.038460	98.557690	98.578200	99.519230	97.630330	12412.000000

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.

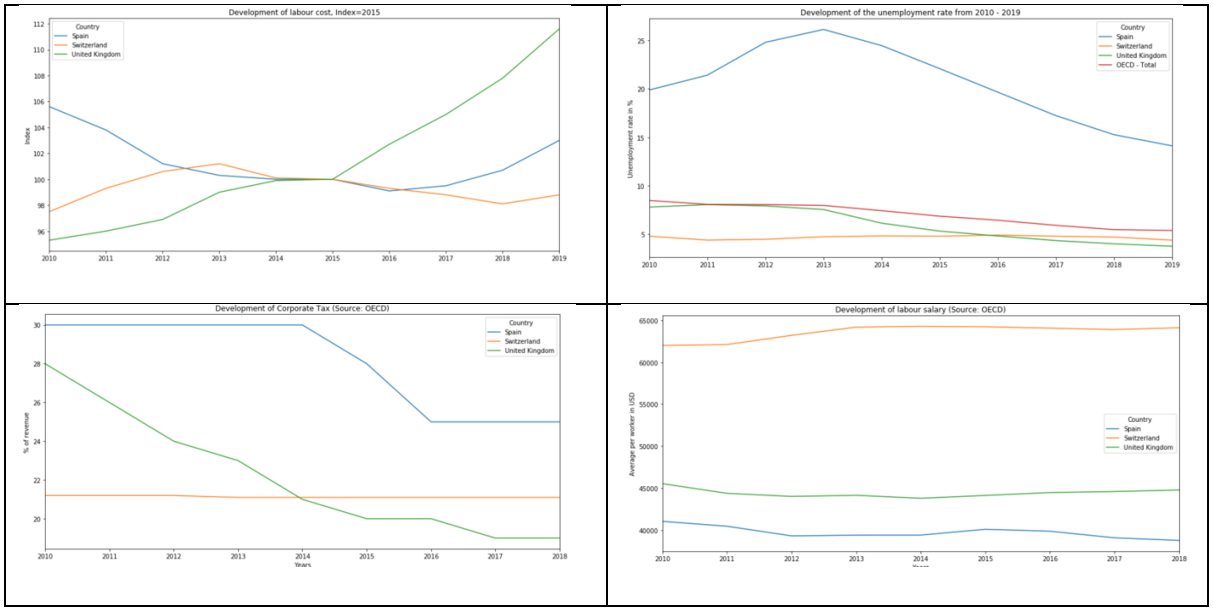
1. Select a country among the three candidates (SPA, CHE, GBR) based on macro economical data as shown above and run a decision tree analysis on the subject to define the key selection points i.e. the high entropy gains.
2. Based on the preselection, and with the data of mainly Foursquare, the different 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:

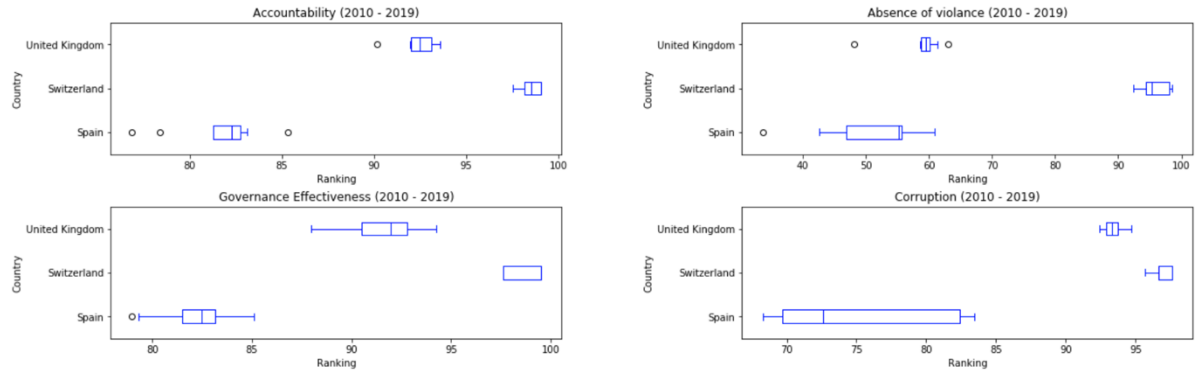


Interpretation of the data

- 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.



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.

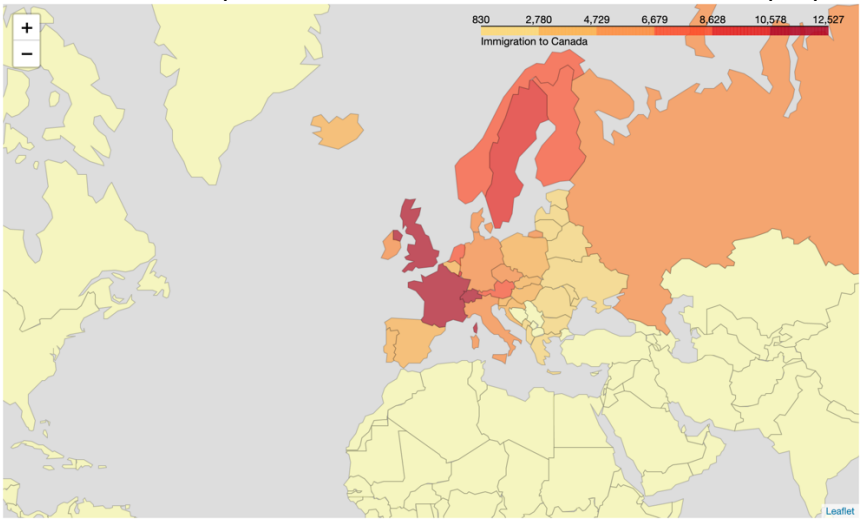
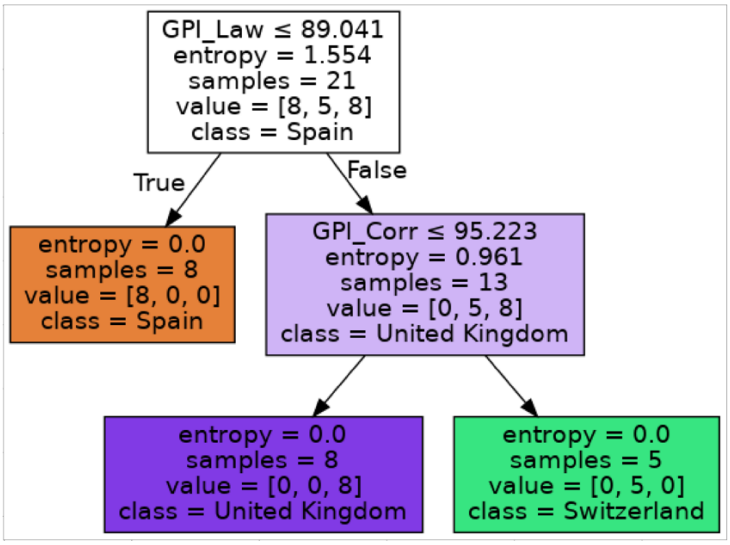


Figure 4: Property cost per m2 in Europe

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 highest 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:

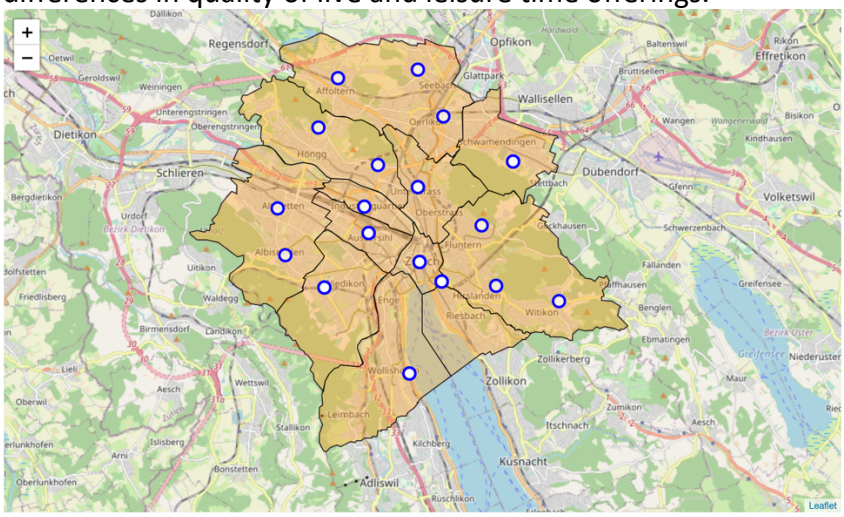
1. Index for “Rule of law”
2. Corruption reliance index



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.



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:

	Neighborhood	Accessories Store	American Restaurant	Argentinian Restaurant	Art Museum	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	Australian Restaurant	Automotive Shop	BBQ Joint	Bakery	Bar	Beer Garden	Bistro
0	Affoltern	0.0	0.0	0.00	0.0	0.00	0.000000	0.090909	0.0	0.0	0.0	0.000000	0.00	0.0	0.0
1	Albisrieden	0.0	0.0	0.00	0.0	0.00	0.000000	0.000000	0.0	0.0	0.0	0.000000	0.00	0.0	0.0
2	Alt-Wiedikon	0.0	0.0	0.00	0.0	0.00	0.090909	0.000000	0.0	0.0	0.0	0.000000	0.00	0.0	0.0
3	Altstetten	0.0	0.0	0.00	0.0	0.00	0.058824	0.000000	0.0	0.0	0.0	0.058824	0.00	0.0	0.0
4	City	0.0	0.0	0.01	0.0	0.01	0.000000	0.000000	0.0	0.0	0.0	0.010000	0.07	0.0	0.0

As mentioned before, each borough has its specialties and characters. Attached the first five boroughs and their 10 most common venues.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Affoltern	Bus Station	Supermarket	Hotel	Miscellaneous Shop	Department Store	Italian Restaurant	Train Station	Athletics & Sports	Diner	Falafel Restaurant
1	Albisrieden	Trattoria/Osteria	Tram Station	Grocery Store	Supermarket	Swiss Restaurant	Yoga Studio	Discount Store	Fast Food Restaurant	Farmers Market	Falafel Restaurant
2	Alt-Wiedikon	Bus Station	Skating Rink	Motorcycle Shop	Hotel	Pool	Tram Station	Asian Restaurant	Supermarket	Light Rail Station	Eastern European Restaurant
3	Altstetten	Supermarket	Swiss Restaurant	Bakery	Plaza	Mediterranean Restaurant	Fast Food Restaurant	Japanese Restaurant	Italian Restaurant	Pool	Asian Restaurant
4	City	Swiss Restaurant	Café	Bar	French Restaurant	Restaurant	Hotel	Cocktail Bar	Italian Restaurant	Department Store	Pedestrian Plaza

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.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
17	Fluntern	Tram Station	Hotel Bar	Fondue Restaurant	Hotel	Cupcake Shop	Deli / Bodega	Department Store	Design Studio	Dessert Shop	Food & Drink Shop
31	Saatlen	Tram Station	Swiss Restaurant	Pizza Place	Athletics & Sports	Yoga Studio	Fast Food Restaurant	Farmers Market	Falafel Restaurant	Electronics Store	Eastern European Restaurant
32	Schwamendingen Mitte	Tram Station	Swiss Restaurant	Pizza Place	Athletics & Sports	Yoga Studio	Fast Food Restaurant	Farmers Market	Falafel Restaurant	Electronics Store	Eastern European Restaurant
33	Hirzenbach	Tram Station	Swiss Restaurant	Pizza Place	Athletics & Sports	Yoga Studio	Fast Food Restaurant	Farmers Market	Falafel Restaurant	Electronics Store	Eastern European Restaurant

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 do. 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.