CAPSTONE: THE MILLENNIAL ENTREPENEUR

ABSTRACT

The Foody Entrepreneur has saved enough money to open a small establishment in Europe.

Laruchelle de Almeida-Bekker

Capstone: Data Science

1. Introduction

1.1 Background

The world is seeing a huge displacement of the Generation Y workforce. The workforce is predominantly becoming occupied by millennials and as studies have shown, millennials differ very much from the Baby Boomer generation in the fact that they are not content with having one job for the rest of their life. The last twenties years have seen a huge increase in entrepreneurs creating their own businesses or goods. As such, technology has also increased, and entrepreneurs don't have to just set up shop at a random location or design something and hope it is what the market needs. With the data available today, entrepreneurs can do their research before-hand to give them a higher probability of success. For this specific problem an EU millennial entrepreneur will be used. As we all know, millennials are very concerned about the environment, so "green" status we also be a determining factor.

1.2 Business Problem

To give the majority workforce a higher probability of success, specifically the entrepreneurial workforce, we the data scientists can assist them in choosing the right location to start their small business. As stated in the background, the business problem we face is to find not only the right location for the entrepreneur to set up shop, but what type of shop will most likely lead to success. This project aims to find an environmentally friendly location for the entrepreneur to set-up shop and give a list of the most shops/venues in the area, to increase his success.

1.3 Interest

Millennials who wish to move to another city, which is more environmentally friendly may find this project interesting, as well as entrepreneurs wishing to start their own small businesses.

2. Data Acquisition and Cleaning

2.1 Data Sources

To find the most environmentally friendly EU country, OECD data will be used. Specifically, CO1 emission released. Secondly, employment rate in the EU will be analysed to find a good fit. The intersection of the most environmentally friendly and highest employment rate country will then be used.

Once the EU country has been found, we will use the biggest cities in the country to plot a map. From there Foursquare will be utilised to find the most common venues in each city to determine what type of establishment should be opened.

2.2 Resources

Data	Data Source	URL
CO2 Emissions	OECD	https://www.oecd-ilibrary.org/energy/data/iea-co2-emissions-from-fuel-combustion-statistics/indicators-for-co2-emissions_data-00433-en
Employment Rate	OECD	https://data.oecd.org/emp/employment-rate.htm
Main Cities	Britannica	https://www.britannica.com/topic/list-of-cities-and-towns-in-Sweden-2050563
Location of Main Cities	LatLong	https://www.latlong.net/category/cities-215-15.html
Venues in Cities	Foursquare	http://www.Foursquare.com

2.3 Data

Below is a screenshot of the CO2 emissions in Europe, the full dataset can be found in the link provided above.

LOCATION	INDICATOR	SUBJECT	MEASURE	FREQUENCY	TIME	Value
AUT	POLLUTIONE	EXPOS2PM25	MICGRCUBM	A	2000	15.46764
AUT	POLLUTIONE	EXPOS2PM25	MICGRCUBM	A	2005	15.70761
AUT	POLLUTIONE	EXPOS2PM25	MICGRCUBM	A	2010	15.36477
AUT	POLLUTIONE	EXPOS2PM25	MICGRCUBM	A	2011	15.67401
AUT	POLLUTIONE	EXPOS2PM25	MICGRCUBM	A	2012	14.50341
AUT	POLLUTIONE	EXPOS2PM25	MICGRCUBM	A	2013	14.28511
AUT	POLLUTIONE	EXPOS2PM25	MICGRCUBM	A	2014	13.53498
AUT	POLLUTIONE	EXPOS2PM25	MICGRCUBM	A	2015	13.59913
AUT	POLLUTIONE	EXPOS2PM25	MICGRCUBM	A	2016	12.65744
AUT	POLLUTIONE	EXPOS2PM25	MICGRCUBM	A	2017	12.68774
BEL	POLLUTIONE	EXPOS2PM25	MICGRCUBM	A	2000	15.97944

Features Kept	Dropped Features	Reason
Location, Indicator & Value	Subject, Measure, Freq and Time	Dropped features not required to find optimal country.
Venues & Location	Tips & Ratings	Tips & Ratings not required to find most common establishment.

3. Exploratory Data Analysis

3.1 CO2 Emissions

LOCATION	INDICATOR	SUBJECT	MEASURE	FREQUENCY	TIME	Value
AUT	EMP	TOT	PC_WKGPOP	A	2015	71.1
AUT	EMP	TOT	PC_WKGPOP	A	2016	71.55
AUT	EMP	TOT	PC_WKGPOP	A	2017	72.2
AUT	EMP	TOT	PC_WKGPOP	A	2018	73.025
AUT	EMP	TOT	PC_WKGPOP	A	2019	73.525
BEL	EMP	TOT	PC_WKGPOP	A	2015	61.8
BEL	EMP	TOT	PC_WKGPOP	A	2016	62.3
BEL	EMP	TOT	PC_WKGPOP	A	2017	63.125
BEL	EMP	TOT	PC_WKGPOP	A	2018	64.45
BEL	EMP	TOT	PC_WKGPOP	A	2019	65.3
CZE	EMP	TOT	PC_WKGPOP	A	2015	70.225
CZE	EMP	TOT	PC_WKGPOP	A	2016	71.95
CZE	EMP	TOT	PC_WKGPOP	A	2017	73.625
CZE	EMP	TOT	PC_WKGPOP	A	2018	74.825
CZE	EMP	TOT	PC_WKGPOP	A	2019	75.125
DNK	EMP	TOT	PC_WKGPOP	A	2015	71.975
DNK	EMP	TOT	PC_WKGPOP	A	2016	72.675
DNK	EMP	TOT	PC_WKGPOP	A	2017	73.225

Figure 1 - CO2 Emissions Table (screenshot)

Using the CO2 emissions table from the OECD page and utilising Tableau Public, the below bubble chart shows EU countries CO2 emissions.

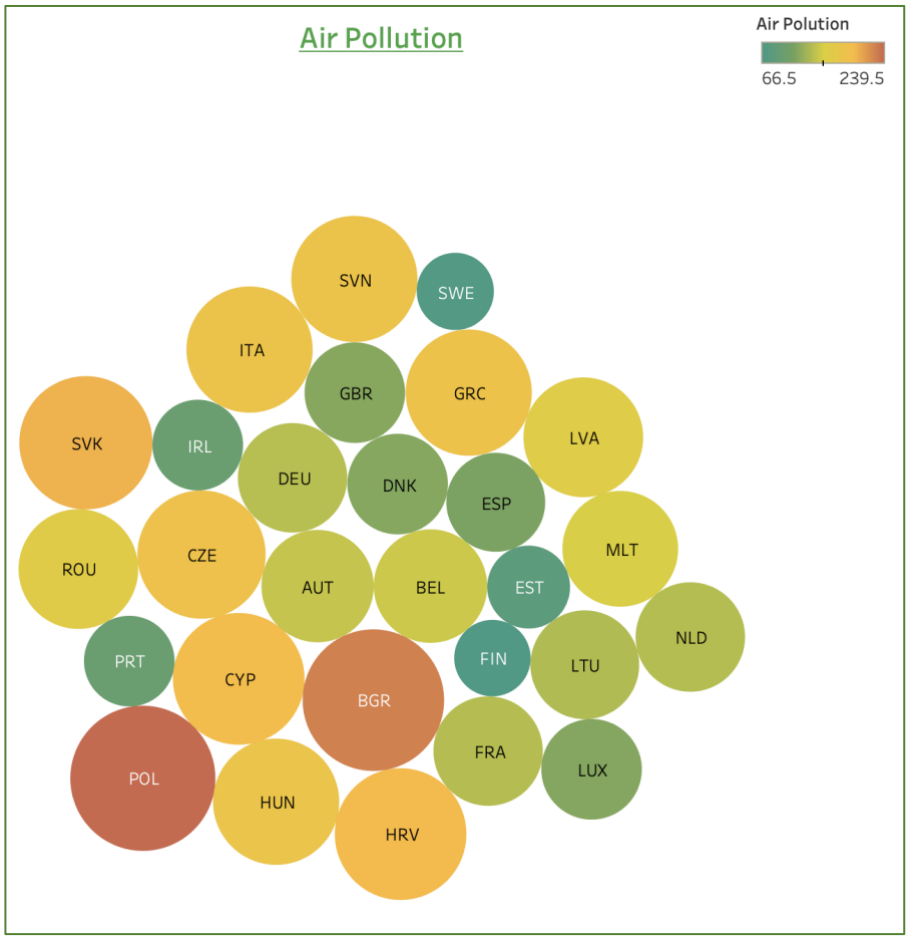

Figure 2 - CO2 Emissions EU

From the bubble chart we can clearly see the two countries with the lowest CO2 emissions are Sweden and Finland.

3.2 Employment Rate

Next the Employment rate is analysed.

LOCATION	Value
POL	27.10558
POL	26.15214
POL	26.00973
POL	25.92949
BGR	25.44274
BGR	25.33827
BGR	24.3186
POL	24.29708
BGR	23.62288
BGR	23.3744
POL	23.22225
POL	22.69148

Figure 3 -Employment Rate of EU countries (screenshot)

EU employment rate per country is graphically illustrated using a Tree Map in Figure 4. The two countries with the highest employment rate are Netherlands and Sweden.

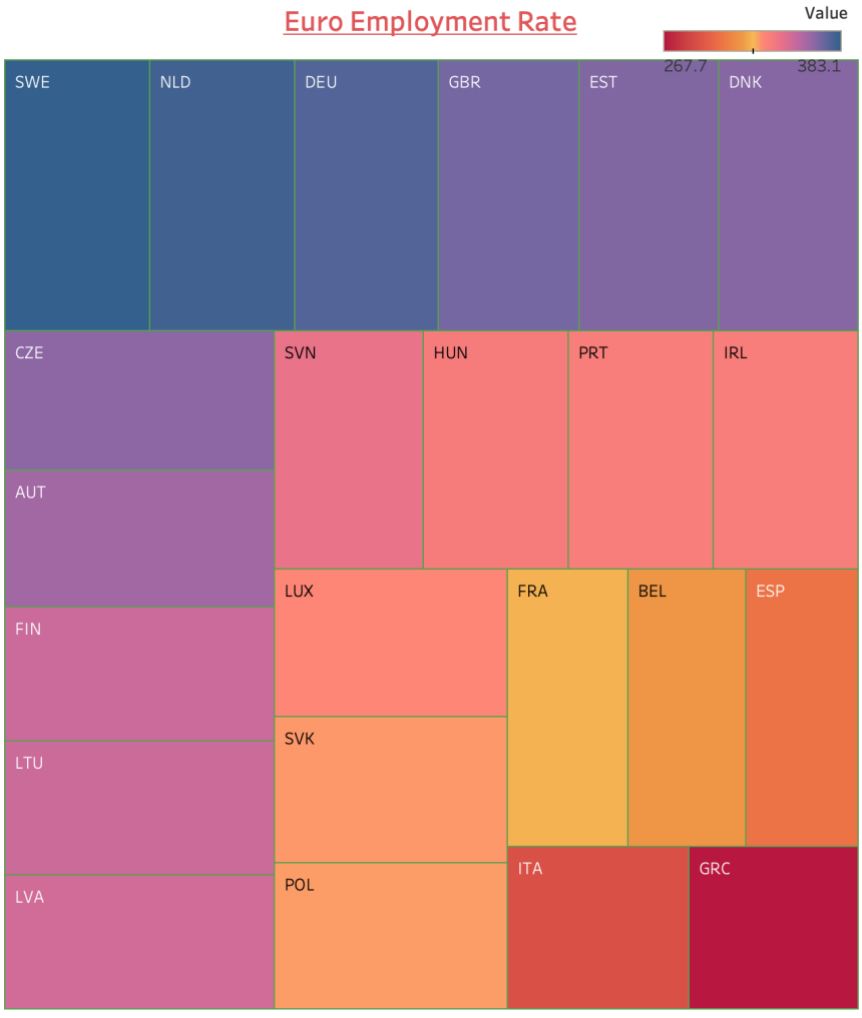

Figure 4 - Employment Rate

3.3 EU Country

Using Figure 2 and Figure 4, it can be clearly seen that Sweden is the country that satisfies the first two problems the millennial entrepreneur has.

Next, the Folium package is used to create a map with Sweden's main cities.

City	Latitude	Longitude
Lulea	65.584816	22.156704
Trollhattan	58.283489	12.285821
Vasteras	59.611366	16.545025
Umea	63.825848	20.263035
Norrkoping	58.588455	16.188313
Stockholm	59.334591	18.06324
Uddevalla	58.351307	11.885834
Vastervik	57.751442	16.628838

Figure 5 - Sweden's Main Cities

3.4 Foursquare Venues

Now that Sweden's main cities are represented by using Longitude and Latitude, Foursquare credentials are used to fetch all venues surrounding the city (limited to 100 venues) with a radius of 500.

When venues in each city have been found, One Hot Encoding will be used to extract all dummy variables necessary to implement K-Means Clustering.

	City	American Restaurant	Arcade	Art Gallery	Arts & Crafts Store	Arts & Entertainment	Asian Restaurant	BBQ Joint	Bakery	Bar	Basketball Stadium	Beer Bar	Bistro	B
0	Lulea	0	0	0	0	0	0	0	0	0	0	0	0	
1	Lulea	0	0	0	0	0	0	0	0	0	0	0	0	
2	Lulea	0	0	0	0	0	0	0	0	0	0	0	0	
3	Lulea	0	0	0	0	0	0	0	0	0	0	0	0	
4	Lulea	0	0	0	0	0	0	0	0	0	0	0	0	

Figure 6 - One Hot Encoding applied

As shown in Figure 6, now we have venues for each major city in Sweden and this will be used to extract the mean of each venue in the city.

```
----Norrkoping----
                venue  freq
0          Restaurant  0.09
1       Shopping Mall  0.07
2               Café  0.07
3  Italian Restaurant  0.07


----Stockholm----
          venue  freq
0         Hotel  0.07
1          Café  0.05
2  Cocktail Bar  0.04
3  Burger Joint  0.04


----Trollhattan----
              venue  freq
0  Sushi Restaurant  0.08
1             Café  0.08
2   Thai Restaurant  0.05
3  Tapas Restaurant  0.05
```

Figure 7 - Frequency of Venues

Using Foursquare and frequency of venues, a table can be created showing the 10 most common venues in each city.

	City	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
0	Lulea	Café	Hotel	Fast Food Restaurant
1	Norrkoping	Restaurant	Shopping Mall	Café
2	Stockholm	Hotel	Café	Cocktail Bar
3	Trollhattan	Café	Sushi Restaurant	Pub
4	Uddevalla	Tunnel	Gym	Yoga Studio

Figure 8 - Most Common Venues

3.5 K-Means Clusters

Now that we have the most common venues per city, K-Means Clustering can be applied. In this reason I used unsupervised learning K-means algorithm to cluster the boroughs. K-Means algorithm is one of the most common cluster method of unsupervised learning. A total of 5 cluster (K=5) will be used as it is determined to be the optimal number of clusters.

Before the clusters can be plotted, two tasks have to be completed. Longitude and Latitude of each city has to be appended to the table shown in Figure 8. Secondly, the cluster labels have to change from float to int, to illustrate the clusters in different colours.

	City	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
0	Lulea	65.584816	22.156704	3	Café	Hotel	Fast Food Restaurant
1	Trollhattan	58.283489	12.285821	4	Café	Sushi Restaurant	Pub
2	Vasteras	59.611366	16.545025	1	Café	Restaurant	Hotel
3	Umea	63.825848	20.263035	1	Hotel	Café	Italian Restaurant
4	Norrkoping	58.588455	16.188313	1	Restaurant	Shopping Mall	Café
5	Stockholm	59.334591	18.063240	1	Hotel	Café	Cocktail Bar
6	Uddevalla	58.351307	11.885834	0	Tunnel	Gym	Yoga Studio
7	Vastervik	57.751442	16.628838	2	Flower Shop	Grocery Store	Department Store

Figure 9 - K-means Clustering Table Result

Figure 10 - K-Means Clustering of Venues in Sweden's Cities

Cluster 0 – Red

	City	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
6	Uddevalla	Tunnel	Gym	Yoga Studio

Cluster 1 – Purple

	City	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
2	Vasteras	Café	Restaurant	Hotel
3	Umea	Hotel	Café	Italian Restaurant
4	Norrkoping	Restaurant	Shopping Mall	Café
5	Stockholm	Hotel	Café	Cocktail Bar

Cluster 2 – Orange

	City	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
7	Vastervik	Flower Shop	Grocery Store	Department Store

Cluster 3 – Blue

	City	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
0	Lulea	Café	Hotel	Fast Food Restaurant

Cluster 4 – Green

	City	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
1	Trollhattan	Café	Sushi Restaurant	Pub

4. Conclusion

After analysing all the datasets and clusters from k-means clustering, it can be deduced that it would be ideal for a millennial entrepreneur to set-up shop in Sweden. To increase the entrepreneur chances for success, it is recommended to open a Café in Vasteras, Umea, Norrkoping or Stockholm (Purple Cluster). As it clearly shows that Cafés are quite popular in Vasteras, Umea, Norrkoping and Stockholm, it would be a good first choice to open for the young businessman.

To conclude, it is suggested that the entrepreneur open a Café in Stockholm, as it is the country's capital and receives the most tourists in a year.

5. Future Directions

There are definitely areas for improvement specifically finding out the rental prices to rent a shop as well as analyse the cost of living in each city.