**COFFEE AND ME:**

**IDENTIFYING THE BEST CITY TO OPEN A CAFÉ**

By: Simon Serrano

1. **Introduction – Business Problem**

The problem is as follows: We are contacted by a small coffee shop franchise that wants to open up a new location. As they are adventurous people, they are willing to put this shop anywhere in the world. The primary concern – as is the case with most businesses – is to choose a location that optimizes profits. To figure out where to open this shop, we will have to consider a number of factors, namely:

1. *What places (cities) consume the most coffee?*
2. *What are the population, GDP, and GDP per capita in these cities?*
3. *How saturated is the coffee shop market in these cities?*

I will elaborate on the relevance of each of these items in the following (Data) section. But the main idea is outlined below:

As we don’t have the bandwidth to look into every city in the world, we will look at the capital cities of roughly the top 50 countries in terms of coffee consumption per capita. Then we will see roughly how many coffee shops already exist in the center (5 km radius) of each of these capital cities. This will give us a sense of roughly how many customers a coffee shop in this area would have. Making adjustments for on-premise vs. off-premise consumption of coffee, we will hopefully get a decent sense of how much coffee each shop would sell in a year. It’s quite possible that the Foursquare API won’t give us all destinations that sell coffee, in which case, we will at least be able to comment on the relative amounts of coffee that could be sold in shops in different countries. This is all we really need, as we just want to determine WHERE we can maximize profits and not WHAT that margin might be. For the sake of this analysis, we assume startup and operational costs are the same everywhere, so really this is a revenue optimization problem.

Essentially, once we understand which places present the best opportunity based on the aforementioned points, we will devise a recommendation as to which city our client should choose for their new shop.

1. **Data**

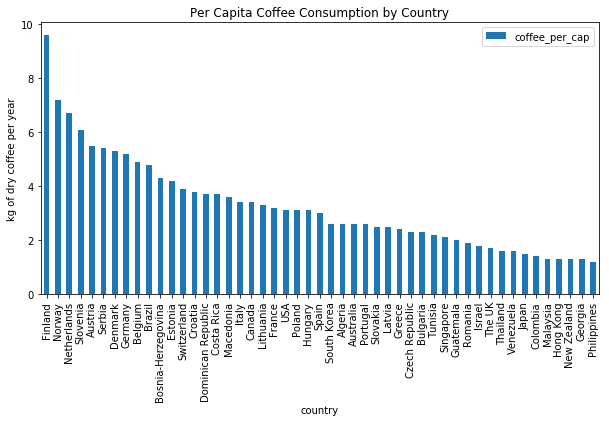
The following factors were referenced in the Introduction – Business Problem section:

1. *What places (cities) consume the most coffee?*
2. *What are the population, GDP, and GDP per capita in these cities?*
3. *How saturated is the coffee shop market in these cities?*

I will now explain the motivation and data sources for these items.

The first – coffee consumption – is obviously something in which our stakeholders would take interest. Specifically, we will use the list provided in the following site for an idea of how popular coffee is (how much is drunk on a per capita basis) in different countries:

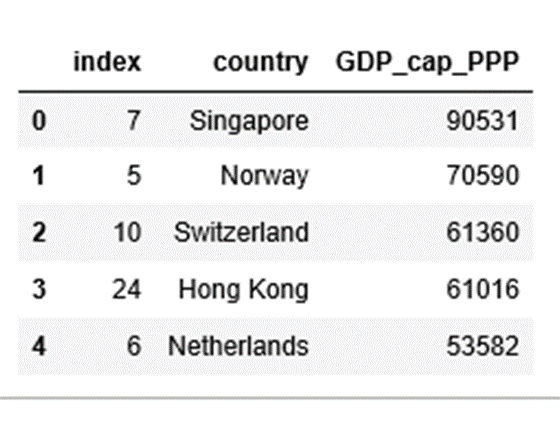
<https://www.caffeineinformer.com/caffeine-what-the-world-drinks>



This is reported at the country level. As we would like to limit our scope to capital cities, we will make the assumption that the per capita consumption provided at the country level is a good reflection of the per capita consumption in the capital city. So, for example, the data source tells us that Serbs consume 5.4kg of dry weight (before brewing process) coffee per capita per year. We will thus assume that coffee consumption is at the same level in the capital city of Belgrade.

The second factor – macroeconomic statistics – will provide us with a metric to help us gauge how wealthy the population in a given country is. Again, we will apply the GDP per capita findings for each country to its capital city. This data will be obtained from the World Bank and can be viewed at the following page:

<https://en.wikipedia.org/wiki/List_of_countries_by_GDP_(PPP)_per_capita>



***Left: Top 5 countries by GDP PPP per capita***

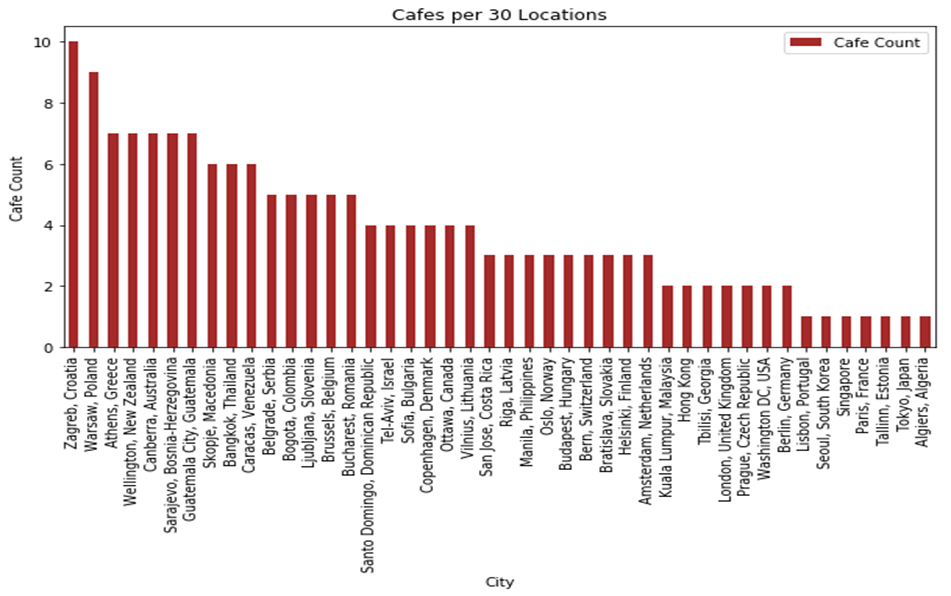
It’s important to note that this is reported in PPP (Purchasing Power Parity) so it already accounts for the differences in value of a dollar in different countries. Population data will be obtained from:

<https://en.wikipedia.org/wiki/List_of_national_capitals_by_population>

Finally, I will use Foursquare to count the number of cafes within a 5 kilometer radius of the center of the capital city (See next page for graphic.). This – in conjunction with population data – will provide us with a population-per-café (PPC) number. Then, we will take this number and multiply it by coffee-consumed-per-capita to understand how much coffee can be attributed to each café. Again, we will use Belgrade as an example: The per-capita-consumption is 5.4kg. The population of Belgrade is about 1.2 million. Once we find the number of cafes (Café Count) in a 5 kilometer radius from the center of Belgrade, we will compute:

*Coffee Potential for new shop (CPNS) = (5.4 kg \* 1,200,000 / Café Count)*

As costs per cup of coffee will certainly vary by country, we will use GDP per capita ratios between countries to adjust for this. And again, these CPNS numbers may be orders of magnitude too high or too low based on the accuracy of location reporting. So we will likely end up relying on the relative CPNS values to determine the city with the best opportunity. The Foursquare results generated the following data on Cafes:



Of particular importance to our business is to consider the relationship between GDP per capita and on-premise coffee consumption. How does the economic situation in a city influence the way in which the local population consumes coffee? Unfortunately we don’t have on vs off-premise coffee consumption data for all of the countries in our data. But we do have *beer* consumption data for the 50 US States:

<https://www.brewersassociation.org/insights/importance-on-premise-craft-brewers/>

<http://scottjanish.com/interactive-map-total-beer-consumption-state/>

And we do have the on/off premise consumption of coffee split in the US in the year 2004:

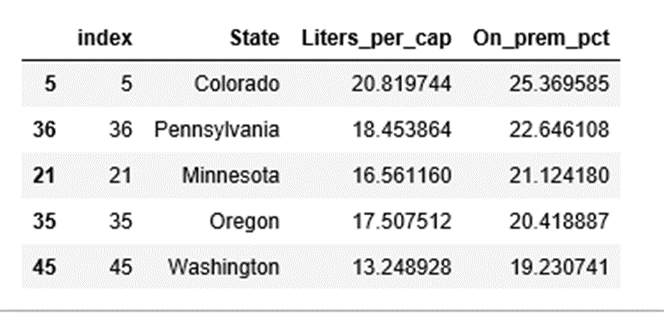
<https://data.worldbank.org/indicator/NY.GDP.PCAP.PP.CD?locations=US>

This will all come in handy. More on this in Methodology!

1. **Methodology**

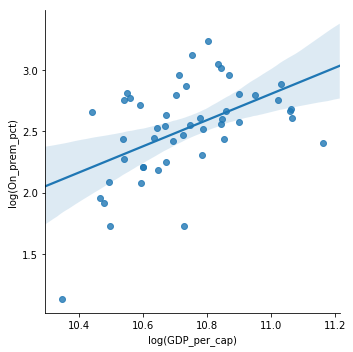
With an understanding of the business problem and the accompanying data, we can refine the methodology that was foreshadowed in the past two sections and detail any exploratory analysis that was done along the way.

With all of the necessary data loaded into pandas DataFrames, we began the coffee shop analysis from an unexpected angle… with beer. As alluded to in the data section, we needed to identify a way to segment coffee sales into on-premise versus off-premise consumption. In case this terminology is foreign to the reader, *on-premise* refers to consumption at a particular venue. So Starbucks or a traditional mom-and-pop coffee shop would be examples of this in the coffee market. *Off-premise*, in contrast, refers to sales from supermarkets or – in the case of beer, for example – liquor stores. Typically, off-premise consumption is more common as it’s cheaper. Unfortunately, we were unable to find on versus off premise consumption of coffee by country. However, we did have this data for beer in the 50 US states.



***Left: Top 5 US States by % of beer purchased on-premise***

So, the first part of this analysis was to use the American beer market as a proxy for the global coffee market. Specifically, we ran a regression of on-premise beer sales as a function of income per capita in each state. After plotting the raw data and the log-transformed data (in both variables), we concluded that the latter gave a better linear representation of its respective points. The former looked to have high heteroscedasticity. Here is the plot:



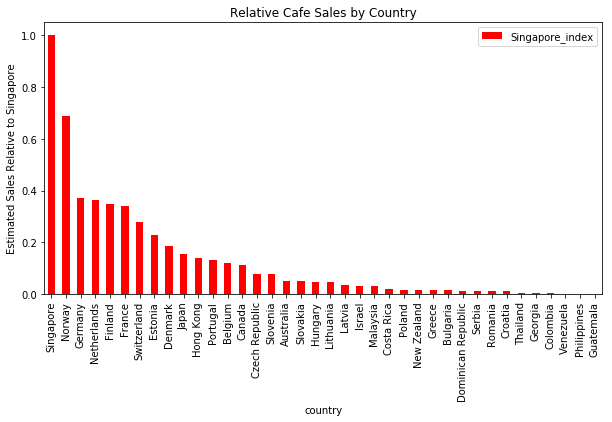
With this regression line – which had an R squared of 0.98 – we were then able to represent how much consumption switches from off premise to on premise as income per capita rises. The result was that a 1% increase in income per capita roughly translated to a 1.1% change in the on-premise fraction (on premise / total consumption). Having found that the on-off split for coffee in the US in 2004 (when GDP per capita was roughly 42,000 USD) was 25%-75%, we were able to use the slope of the regression line to calculate the approximate on-off split in the countries we were analyzing. We did this by finding the difference in GDP per capita between that country and the US’s 2004 number and then applying the slope of the regression line from the beer regression to compute the deviation from the 25% anchor.

Having established our beer proxy as a reasonable way of understanding conversion from off to on premise, we then moved on to the coffee datasets themselves. This part represents a majority of the work in the IPython file, but the reasoning is fairly straightforward. We were able to find online the top 49 countries for per-capita coffee consumption. Obviously, if we are to identify a good market for our coffee shop, we want to know what countries consume a lot of coffee. Then, we used Foursquare API location data to pick out venues within a 5-kilometer radius of the center of the capital city in each of these countries. Upon looking at the venue categories, we combined all ones that we believed appropriately resembled cafes. Some were called ‘coffee shops,’ ‘donut shops,’ ‘kafenios,’ etc. From here we got a count of such places within that radius in each country. Unfortunately, the Foursquare results only give a sampling of locations within the specified radius. So we could not – as was hoped before – establish a café density in each city. Instead what we did was make the following assumption: Locations are directly proportional to population. That is to say, 30 locations (not necessarily cafes) in London will represent the same absolute population as 30 locations in Lisbon and so on. This precludes us from needing to use the population densities of each country. What we then did was take the count of coffee shops (out of the 30 locations) and multiply by the per-capita-coffee consumption in each corresponding country and then multiply this by the on-premise fraction discussed earlier. This represents the *relative* coffee consumed in each café in like-populated areas around the world. But as our focus is on profit or in our case – taking startup cost as constant across markets – revenue, then we wanted to adjust for differences in per-capita income across countries. So we divided each nation’s per-capita GDP by the maximum value and multiplied this index by our ‘coffee consumed in each café’ figure to approximate a relative revenue-per-café in each country. Sorting this figure gave us the most profitable markets – the best places to start a coffee shop!

Finally, we wanted to identify other markets that we believed were trending in line with the leading market from our analysis. By ‘trending’ we are referring to how similar their venue frequencies were from the results of the Foursquare data retrieval. To do this, we utilized the K-means clustering algorithm. The result is of course not the same every time it’s run due to the algorithm sometimes finding local – as opposed to – global extrema.

1. **Results**

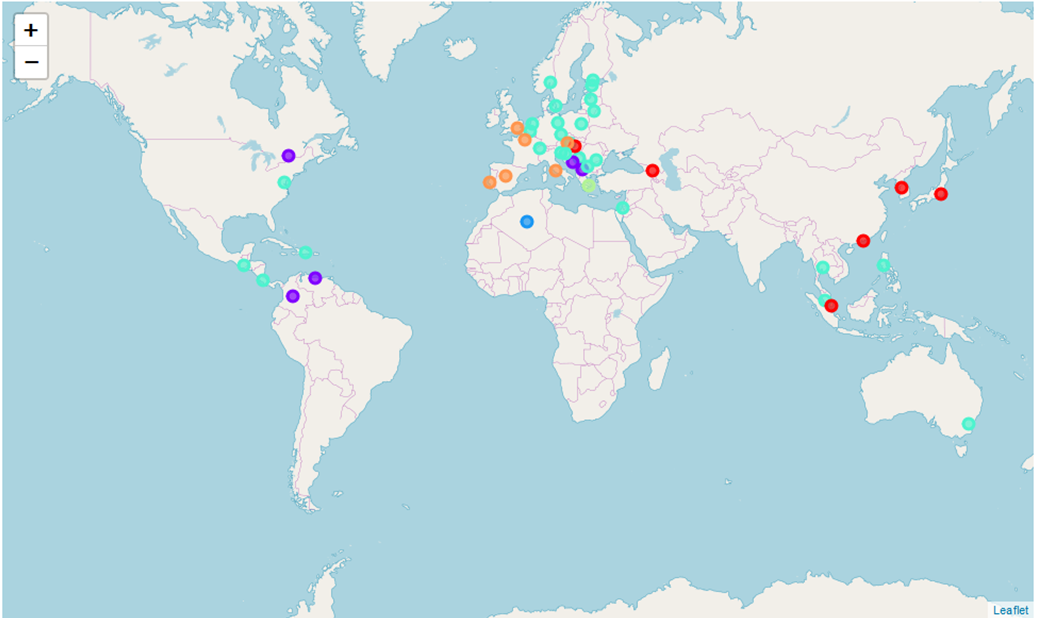
After all of this analysis, it was concluded that Singapore was the best market to open a coffee shop based on the revenue-optimizing criterion stated at the beginning of this report. Specifically, the relative revenues suggested that not only was Singapore the best place to open the coffee shop, but it was estimated to be over 1.4 times as profitable as the second best market – Norway, which had a revenue index of 0.69 compared to the 1.00 of Singapore. Germany was in third with an index of 0.37.



***Left: Results of relative sales analysis***

Unfortunately, Spain, Italy, Austria, and Algeria were removed from the analysis either due to 0 cafes being located (which was particularly dubious for well-established coffee-cultures like Italy and Austria) or simply not enough venues being located (2 for Algeria compared with 30 for all others).

For the cluster analysis, it was found that Hungary, Hong Kong, Japan, South Korea, and Georgia were in Singapore’s cluster. This diverse mix of countries represent places that most closely resemble Singapore in terms of venue category frequencies and thus could be interpreted as trending in line with Singapore. Perhaps this will result in their also being favorable areas to open coffee shops in the future!



***Left: Map of clusters from similar-venue comparison***

1. **Discussion**

Aside from Singapore, the best markets for opening a coffee shop were dominated by Northern Europe. Places like Norway, Germany, the Netherlands, and Finland were high up on our list of favorable markets. What’s particularly interesting about that is that these places have very high per-capita consumption of coffee as well as high income per-capita. The fact that they’re so high on the list implies that supply of coffee shops has yet to reach demand for on-premise coffee as economies improve and consumption trends toward on-premise purchase. This of course assumes that the Foursquare data provides an accurate sampling that would enable appropriate frequency-of-café calculations.

In that same vein, we’d be remiss were we not to comment on the uniqueness of a Southeast Asian nation topping the list over established coffee countries. This reminds us of the value of capturing trends in consumption. Singapore is an extremely wealthy country. From a consumer goods perspective, successfully pushing a product that isn’t already popular in such a wealthy market is sure to make stakeholders happy!

1. **Conclusion**

At the risk of repeating ourselves, we once again note the utility of having tools such as Foursquare to extract sample data from any region of the world. This was indispensable in understanding the prevalence of cafes in various countries around the world. From here we could incorporate what we know about average coffee consumption in different countries as well as the general trend of evolution from off-premise to on-premise purchases as economies grow to make a relative comparison of expected coffee sales in a café in each country. This sort of analysis is of course not unique to coffee nor is it unique to the food and beverages industry as a whole. Perhaps the biggest question that would merit further investigation is: How well does the Foursquare data actually represent frequencies of locations given that it appears to give seemingly random samples. With more time and capacity, this location extraction could be done iteratively and then the results averaged to ensure reasonable means were obtained. Nevertheless, this surely was a worthwhile and fascinating excursion into the world of Python and coffee. Thanks for reading!