# Data Science Capstone Project Report

## Relocating From Athens To Stockholm

Eric de Connick

October 26, 2020

## Introduction

## Understanding

My client has a company that helps people relocate from a country to Stockholm, Sweden. My client lives in Stockholm most of the time and has a Greek wife stationed in Athens. He has learned from his wife that, due to the current economic crisis in Greece, many young professionals are eager to leave Athens and relocate to Stockholm. These young professionals are fond of their current neighbourhoods in Athens and what they offer in terms of venues and would probably like to move to similar neighbourhoods in Stockholm.

## Business Problem

My client would like to know if it is possible to identify which neighbourhoods in Stockholm are similar to particular neighbourhoods in Athens?

## Interest

The stakeholders of the project are my client and the professionals that would like to relocate from Stockholm to Athens.

## Data

## Data Requirements

I intend to solve this problem by segmenting and clustering location data from Foursquare for neighbourhoods in Athens and Stockholm to identify neighbourhoods in Athens and Stockholm that belong to the same cluster.

The data from Foursquare that I intend to use is data relating to venue categories (gyms, shops, restaurants, parks etc) for the neighbourhoods in question. For instance using the Foursquare API it is possible to get a list of all venues along with their categories for each neighbourhood. Creating a data frame where each row contains a neighbourhood and the frequency of each venue category, I hope to be able to cluster the neighbourhoods based on this data.

## Source

I will also be using data from <https://en.wikipedia.org/wiki/Category:Districts_of_Stockholm>

and <https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Athens> to construct a data frame with names, latitude and longitude coordinates for neighbourhoods in Stockholm and Athens respectively. The latitude and longitude coordinates for each neighbourhood will be used to retrieve the location data above from Foursquare.

## Data Preparation

The data was prepared by creating an Excel spreadsheet with two columns City and Neighbourhood from the data sources mentioned above (I was not able to scrape the web pages).

The spreadsheet was then loaded into a data frame and latitude and longitude columns were added for each city neighbourhood.

Using the latitude and longitude coordinates for each neighbourhood and the Foursquare API, information about each all the venues within a 500 meter radius from the centre of each neighbourhood were added to the data.

This data was then transformed, using one-hot encoding, into a data frame containing all the mean frequencies for all venue categories for each neighbourhood.

## Data Cleaning

While preparing the data, it became clear that latitude and longitude coordinates could not be found for some Athens neighbourhoods.

This problem was rectified by using the corresponding Greek (letters) names instead of the English names found on the website over Athens municipalities.

## Methodology

To answer the question about which neighbourhoods in Athens are similar to which neighbourhoods in Stockholm the following tasks were identified:

* Create a data frame with the mean frequency for each venue category in a neighbourhood (see section Data Preparation)
* Use k-means clustering to cluster the neighbourhoods based on the frequencies of the different neighbourhood categories.

## Exploratory Data Analysis

Having a data frame with the longitude and latitude coordinates of each neighbourhood, the data was visualized by displaying a map for each city with the neighbourhoods superimposed on the respective maps.

From the visualization of Athens and its neighbourhoods it was discovered that there was an outlier which did not seem to be in Athens at all.

The Data Cleaning step of the process had to be re-visited and this issue was also rectified by correcting the latitude and longitude coordinates by using the Greek name of the neighbourhood.