IBM

PROFESSIONAL CERTIFICATE

Data Science

Powered By coursera

Tourism in Chile and Peru post COVID-19: clustering touristic neighborhoods for boosting countries’ economies

Author: Diego Ignacio López Orellana

Date: August 27th 2020

2. Description of the Data

2.1 Datasets used in the Project

In this project, we will gather reliable data and preprocess it using different techniques to finally get the boroughs, neighborhoods, coordinates and the most common venues, in Lima and Santiago de Chile, to be analyzed (mentioned in the Business Problem). The sources from where we will obtain our datasets and the information that each one contains, according to each city, are the following:

2.1.1 Lima, Peru

For Peru's capital, we work with the boroughs and neighborhoods from Lima and Callao, as both provinces are merged in the urban Lima Metropolitan Area **[10]**. In the case of its neighborhoods (known in Peru as *'barrios'*), we work with the *'centros poblados'* as they are the smallest political-administrative circumscriptions of the country; due to the lack of data about Lima's neighborhoods.

The information about Lima's *'centros poblados'* is gathered from the *'Plataforma Nacional de Datos Abiertos'* webpage **[11]**, which belongs to the Peruvian government. From that page, we get a **.xlsx file** called *'ListadoCentroPobladosMTC.xlsx'*. It contains the following relevant columns from the Peruvian *'centros poblados'*:

Column (Feature)	**Description**
Provincia	Each of the Peruvian provinces. ***Data type: String.***
Distrito	Each of the Peruvian districts. ***Data type: String.***
CCPP	Each Peruvian *'centro poblado'*. ***Data type: String.***

Latitud (coord X)	Latitude coordinate of each Peruvian *'centro poblado'*. ***Data type: Float.***
Longitud (coord Y)	Latitude coordinate of each Peruvian *'centro poblado'*. ***Data type: Float.***
CLASIFICACIÓN INEI	Represents if each Peruvian *'centro poblado'* is rural or urban. ***Data type: String.***

Table 1: *ListadoCentroPobladosMTC.xlsx* file columns and their descriptions.

2.1.2 Santiago de Chile, Chile

The capital of Chile is divided into 32 boroughs or communes (called *'Comunas'*) **[12]**, which are subdivided into hundreds of neighborhoods, known as *'Barrios'*. For this project, we will also add four boroughs located in Santiago conurbation: Padre Hurtado, Peñaflor, Puente Alto and San Bernardo.

The dataset containing its boroughs will be retrieved from Wikipedia **[13]** by applying Web Scraping to get table from an HTML webpage. Unfortunately, there is no information available about their neighborhoods nor latitude or longitude values. Therefore, we will retrieve both coordinates by using the ArcGIS World Geocoding Service **[14]**, to convert each neighborhood/borough name into their coordinates.

On this operation, we make calls to its database by sending **Request/GET** sentences to retrieve this information. It will be stored in a **.json file**, which will be parsed and converted to a Pandas data frame to start working with its information.

After creating both datasets, we will join them to create a ***Pandas data frame*** containing all the information required. The data frame will contain the following columns:

Column (Feature)	Description
Neighborhood	Name of each neighborhood or *'barrio'* in Santiago. Here, we assume boroughs as neighborhoods. ***Data type: String.***
Location	Position in a cartesian plane with respect to the direction of each neighborhood. ***Data type: String.***
Latitude	Latitude coordinate from each neighborhood in Santiago. ***Data type: Float.***
Longitude	Longitude coordinate from each neighborhood in Santiago. ***Data type: Float.***

Table 2: *Pandas* data frame columns and their descriptions, for Santiago de Chile.

The latitude and longitude coordinates for the neighborhoods in Santiago de Chile, will be retrieved making calls to a database, which will be explained next.

2.2 APIs used to gather venues and their coordinates

As mentioned before, to gather the coordinates of each neighborhood for Santiago de Chile, where this information is missing, we will make calls to the **ArcGIS database by its World Geocoding REST API [14],** to get the latitude and longitude values required per neighborhood.

On the other hand, to cluster the neighborhoods based on their venues, we will use the **Foursquare API [15]** to make calls to its database and retrieve a **.json file** containing the **venues of the different**

neighborhoods of each city within a radius of 2500 meters for Lima, and 7000 meters for Santiago de Chile, to determine which ones are the most common based on their categories.

2.3 Python libraries used for the Capstone

To import the data, preprocess, make an exploratory analysis and then model it and evaluate the results, we need to import several Python libraries to perform these several tasks that we will carry out during the final project. This is because these libraries contain the functions and methods needed to perform the segmentation of the neighborhoods. The following table summarizes each Python library that will be used in the Final Capstone and its description:

Python Library	**Description**
NumPy	Math library to work with N-dimensional arrays.
Pandas	Library for importing, manipulating and analyzing data in data frames.
JSON	Contain methods to handle JSON files.
GeoPy	Python client which contains geocoder class for several geocoding services to retrieve coordinates of different places.
BeautifulSoup 4	Popular Python library, widely used for Web Scraping: it allows to parse and pull data out of HTML and XML files.
Requests	Library to handle HTTP requests.
Matplotlib	Popular Python plotting package with contains several modules for 2D and 3D plotting.
Scikit Learn	Free Machine Learning library to work with several ML algorithms, performs most of the tasks in a ML pipeline.
Folium	Data visualization library, it is used to visualize geospatial data by the creation of maps at any location in the world.

Table 3: Python libraries used in the Capstone and their descriptions.

3. Methodology

This project is based on a structured Data Science working methodology involving a series of steps from Business and Data understanding to sketch the suggestions from the outcomes of the project to its potential stakeholders: **Chilean and Peruvian touristic agencies, the Governments and their Tourism Ministries from Chile and Peru**.

The following image [16] briefly resume these steps:

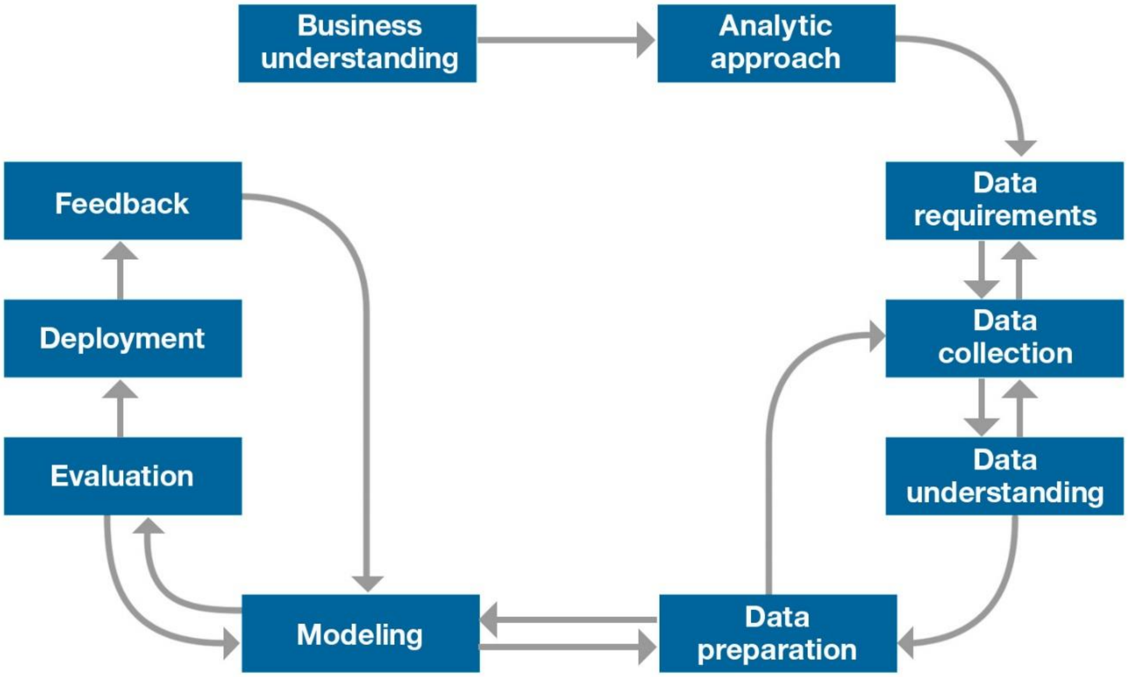

Figure 1: Data Science Methodology used for the project. Source: IBM Big Data and Analytics Hub.

We can summarize the tasks to perform in this project as follows:

1. Business and Data understanding to come up with a problem and define the data requirements to solve it.
2. Data Preparation: wrangling, formatting and preprocessing it to prepare data for further analysis.

3. Exploratory Data Analysis: this step is done to summarize the main venues from each neighborhood to later group them based on the categories of their most common venues.
4. Modelling: with the datasets from Lima and Santiago de Chile, we apply the k-Means Clustering Algorithm per each city to cluster their neighborhoods. This is done to find which ones are more similar based on their venues' categories.
5. Evaluation: each model is evaluated to segment and label which neighborhoods are the most touristic ones.
6. Deployment: write and show to the potential stakeholders the touristic neighborhoods and governments of Chile and Peru, to improve their touristic packages post COVID-19 pandemic.

After this step, the data frames from the touristic neighborhoods of both cities are merged into one. Then, this data frame is used in the following steps:

7. Re-Modelling: the data frame which groups the touristic neighborhood is used to run again the k-Means Clustering Algorithm to cluster and find which neighborhoods are more similar based on their most popular venues' categories.
8. Re-Evaluation: from this model, we segment and label the touristic neighborhoods across the cities studied: Lima & Santiago de Chile.
9. Re-Deployment: again, we deliver a stack of proposals to the touristic neighborhoods and governments from Lima and Santiago de Chile. But now, the focus is to propose them cooperative alliances and policies that can help their countries to recover the touristic industry, in a faster and more effectively way, their economic recovering.

Bibliography

[10] Wikipedia. (2020, July 18). *Lima metropolitan area*. Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Lima_metropolitan_area

[11] OTI - Ministerio de Transportes y Telecomunicaciones Perú. (2018, March 23). *MTC - Centros Poblados*. Retrieved from Plataforma Nacional de Datos Abiertos - Gobierno de Perú: https://www.datosabiertos.gob.pe/dataset/mtc-centros-poblados

[12] Wikipedia. (2019, December 29). *Santiago Province, Chile.* Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Santiago_Province,_Chile

[13] Wikipedia. (2020, August 22). *Anexo:Comunas de Santiago de Chile*. Retrieved from Wikipedia: https://es.wikipedia.org/wiki/Anexo:Comunas_de_Santiago_de_Chile

[14] ArcGIS for Developers. (2020, January 01). *Geocoding with ArcGIS*. Retrieved from ArcGIS for Developers: https://developers.arcgis.com/features/geocoding

[15] Foursquare. (2020, January 1). *Foursquare Developers - Endpoints*. Retrieved from Foursquare: https://developer.foursquare.com/docs/places-api/endpoints/

[16] Rollins, J. (2015, August 24). *Blogs - Why we need a methodology for Data Science*. Retrieved from IBM Big Data and Analytics Hub: https://www.ibmbigdatahub.com/blog/why-we-need-methodology-data-science