Data Science Capstone Project: Correlation between number of traffic accidents per neighborhood and the surrounding venues in Mexico City

Hugo Rocha
IBM-Coursera, Data Science Specialization Final Project

June 18, 2019

Tabla de contenido

Introduction

This report reflects the knowledge acquired during the nine courses of the IBM-Coursera Data Science professional specialization. The final project consists on the application of the acquired skills and the leverage of the Foursquare location data, to explore the neighborhoods of a city of choice and/or to come up with a problem that can be solved by using the Foursquare location data.

In this report Mexico City was selected to try to find if there is a correlation between the venues present close to the place where the traffic accidents are reported and the number of accidents.

According to reports from the "Instituto Nacional de Seguridad Pública" (INSP), Mexico occupies the seventh place in the world and the third place in Latin America in road traffic deaths. Mexico City is the largest city in the country with a population of 8.9 million in 2015, according to data from the "Instituto Nacional de Estadística y Geografía" (INEGI). New insurance companies and new transportation platforms such as Uber and DiDi, need to identify plausible risks when they intend to start operations in new countries and/or cities. A proper risk analysis could help these types of companies to establish prices and define the strategies required to mitigate issues that may arise with the users that are demanding their services. In particular for companies with disruptive business models, such as Uber and DiDi, it is vital for their survival to help drivers to minimize the probability of traffic accidents and to ensure the security of their users.

In order to identify the correlation between the venues present in a neighborhood and the occurrence of traffic accidents, the Foursquare location will be used and a dataset of traffic accidents from 2014 to 2015 in Mexico City has been downloaded in order to carry out the proposed analysis. In the next section a detailed explanation of the used data is presented.

Data Description

A Mexico City's dataset has been chosen as the observation target due to the following reasons:

- The availability of two years of historic data of traffic accidents registered with information of the neighborhood where the accident happened.
- The traffic accident dataset contains information on type of accident, neighborhood and the coordinates of the place where it happened.
- The availability of geographic data that can be used to visualize the dataset onto a map.

The dataset will be composed of the following sources:

- The open dataset of traffic accidents from Mexico's City government. It contains data from 2014 and 2015. It was obtained from this link: https://datos.cdmx.gob.mx/explore/dataset/incidentes-viales-c5/table/?dataChart=eyJxdWVyaWVzIjpbeyJjaGFydHMiOlt7InR5cGUiOiJsaW5lIiwiZnVuYyI6IkFWRyIsInlBeGlzIjoiY29sdW1uXzEiLCJzY2llbnRpZmljRGlzcGxheSI6dHJ1ZSwiY29sb3IiOiIjNjZjMmE1In1dLCJ4QXhpcyI6ImNvbHVtbl8xIiwibWF4cG9pbnRzIjoiIiwidGltZXNjYWxlIjpudWxsLCJzb3J0IjoiIiwiY29uZmlnIjp7ImRhdGFzZXQiOiJpbmNpZGVudGVzLXZpYWxlcy1jNSIsIm9wdGlvbnMiOnt9fX1dLCJkaXNwbGF5TGVnZW5kIjp0cnVlLCJhbGlnbk1vbnRoIjp0cnVlLCJ0aW1lc2NhbGUiOiIifQ%3D%3D
- Foursquare API which will be used to obtain the venues in the vicinity of the place where the accidents happen.

Data preprocessing considerations:

1. The traffic accidents dataset must be formatted:
   a. Drop all rows belonging to year 2014, we will only work with data from 2015.
   b. We only need the following columns:
      i. delegacion_inicio
      ii. incidente_c4
      iii. latitud
      iv. longitud
   c. In order to simplify the marking process, the names of the columns will be changes to their translated version in English as follows:
      i. delegacion_inicio -> neighborhood
      ii. incidente_c4 -> accident
      iii. latitude -> latitude
      iv. longitud -> longitude
   d. The description of the accidents will be changed for their equivalent in English.
2. The resulting dataset will be grouped by neighborhood.
3. For each accident, pass its coordinates to the Foursquare API. The "explore" endpoint will return a list of the surrounding venues in the defined radius.
4. Count the occurrence of each venue type per accident. Then apply one hot encoding to turn each venue type into a column with their occurrence as the value.