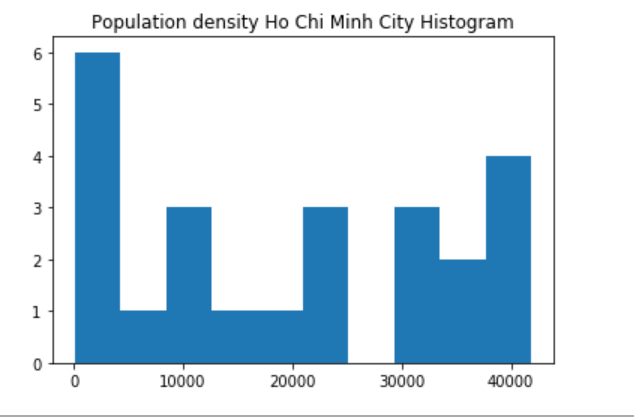
BATTLECAPSTONE PROJECT: THE BATTLE OF THE NEIGHBORHOODS

NGUYEN TAN HUY

March 17, 2020

1. INTRODUCTION
   1. Background  
      Coffee is the most popular drinking in the world. The coffee industry is highly profitable. In Viet Nam, Coffee is the regular drinking. People drink coffee every morning. They are working, dating, playing game in the coffee shop. My company wants to open a new coffee shop in Ho Chi Minh City which is millions of coffee shops. We want to find locations that are not already crowded with the coffee shop in Ho Chi Minh City and many other services as possible.
   2. Problems  
      In Ho Chi Minh City, we have 24 districts with 324 wards. Some districts with high population density, other low population density. And we have millions of coffee shops in these districts, find out where can we open a new coffee shop is a difficult task. With a little help of data, we can have an overview of coffee shop market in Ho Chi Minh City. Find out the best wards to open a new coffee shop.
2. DATA ACQUISITION AND CLEANING
   1. Data Source  
      District Code, District Name, Population, Acreage can be found in [Wikipedia](https://vi.wikipedia.org/wiki/Th%C3%A0nh_ph%E1%BB%91_H%E1%BB%93_Ch%C3%AD_Minh). Ward Code, Ward Name can be found in [official Ho Chi Minh City government website](http://www.pso.hochiminhcity.gov.vn/web/guest/home;jsessionid=110E7DB2FD8D09EF0F14A33EB055CC9E). Data on Wikipedia about population is also getting from official Ho Chi Minh City government website in 2018. New population data is not public yet.
   2. Data collection and data cleaning  
      Data download and scraped from multiple sources. Data from Wikipedia contain population and acreage without ward code and district code, data from official Ho Chi Minh City government website contain ward code and district code. We need to define a list of district code and combine two data frames into one data frame.   
      Getting longitude and latitude from google API. Because of the diverse shapes of the ward, and coordinates getting from API is not in the middle of the ward. I trying to get the exact central point of each ward hand by hand using google map.  
      Getting data from foursquare using radius in foursquare API. Radius will be found be getting average acreage each ward. Here is radius formula  
      After we got longitude and latitude. We get venue location information using the foursquare API. We will all venues with 3 [categories](https://developer.foursquare.com/docs/resources/categories). Coffee shop venues categories which are all kind of coffee shop in Ho Chi Minh City includes Cafeteria, Pet Café, Corporate Coffee Shop, Gaming Café… Long term venues includes: Residence, Home, Residential Building, Office building, government building which are all kind of venues provide the regular customer. Short term venues include all kind of shops, entertainment building… which are all kind of venues provide traveler, the non-regular customer.   
      We will analyze our data base on 2 features: average acreage per venues and average population per venues each district.
3. EXPLORATORY DATA ANALYSIS
   1. Calculation of population density  
      We don’t want our coffee shop open in rural area which is low population density. We will choose districts which are have high or average number of population density.  
        
      Most of the districts have population density higher than 10000 persons per acreage. We will segment our districts into 3 segments low density which is lower than 10000, average density which is between 10000 and 30000 and high density which is higher than 30000 persons per acreage. We only choose districts with average and high population density.
4. DISCUSSION
5. CONCLUSION AND RESULTS