

PHÒNG LẬP TRÌNH & MẠNG

TRUNG TÂM TIN HỌC

ĐẠI HỌC KHOA HỌC TỰ NHIÊN

⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯

Đồ án:

ỨNG DỤNG CLUSTERING VÀ PHÂN TÍCH DỮ LIỆU VÀO VIỆC CHỌN ĐỊA ĐIỂM MỞ QUÁN ĂN VIỆT NAM TẠI TPHCM

Giáo viên hướng dẫn: Nguyễn Quan Liêm

Học viên thực hiện: Trần Hoàng Long

Contents

[1. Lời giới thiệu 3](#_Toc80814068)

[2. Thu Thập và làm sạch Dữ liệu 4](#_Toc80814069)

[2.1 Data sources 4](#_Toc80814070)

[2.2 Data cleaning 4](#_Toc80814071)

[3. Exploratory Data Analysis 9](#_Toc80814072)

[4. Preprocessing 12](#_Toc80814073)

[5. Modelling 15](#_Toc80814074)

[5.1 Elbow Test 15](#_Toc80814075)

[5.2 Clustering 17](#_Toc80814076)

[5.3 Further Analysis 19](#_Toc80814077)

[6. Findings 20](#_Toc80814078)

[6.2 Discussion 20](#_Toc80814079)

[6.2 Conclusion 21](#_Toc80814080)

[6.3 Comments 22](#_Toc80814081)

# Lời giới thiệu

Đồ án này được thực hiện để tìm một địa điểm thích hợp để mở một quán ăn Việt Nam bằng Data Science. Đồ án này hướng đến những người có ý định mở một nhà hàng Việt Nam tại TPHCM nhưng không biết nên mở ở quận nào.

Vì hiện tại số lượng nhà hàng Việt Nam ở TPHCM là nhiều vô kể , vậy nên trong Đồ Án này em sẽ tìm ra những khu vực mà có mật độ cạnh tranh quán ăn Việt Nam thấp nhất để giảm thiểu rủi ro cạnh tranh, ngoài ra em còn xem xét đến những yếu tố khách quan như chi phí bất động sản ( nếu thuê, hoặc mua địa điểm kinh doanh), mật độ dân số ( chẳng ai đi mở quán ăn kinh doanh ở nơi không có người ) và mật độ công ty / nhà xưởng /... ( để bán cho công nhân đến ăn sáng / ăn trưa/ ăn tối hoặc ghé quán để " lai rai". Tóm lại là sẽ phân tích và củng cố những findings của mình một cách thật là data driven.

Ngoài ra, đồ án này còn có thể được dùng làm cảm hứng ý tưởng cho nhiều loại hình kinh doanh khác nữa chứ không giới hạn ở việc mở một Quán Ăn Việt Nam.

# Thu Thập và làm sạch Dữ liệu

## Data sources

Những data cần thu thập là :

- Số quận huyện của TPHCM, cùng với kinh độ vĩ độ ( để vẽ lên bản đồ)

- Mật Độ dân số , số công ty , nhà máy , xí nghiệp của từng quận.

- Mật độ những địa điểm của từng quận huyện của TPHCM ( sẽ sử dụng Google Places API và Foursquare API )

## 2.2 Data cleaning

Table

Description automatically generatedNhững dữ liệu được thu thập về , ngoại trừ dữ liệu từ API Foursquare tất cả đều cần phải được ETL ( đối với dữ liệu dạng JSON của Google Places API) hoặc process và cleansing để có thể đưa về dạng Dataframe.

Việc clean data có thể gồm nhiều task và nhiều bước, gồm cả những việc đơn giản như tách khoảng trống trong chuỗi cho tới những việc đặc thù như chuyển đổi tiếng Việt từ có dấu thành không dấu để thư viện Geocoder có thể hiểu được. Phần lớn việc clean data và chuẩn hóa dữ liệu đến từ việc phải lọc lại các dữ liệu trong content được crawl về từ beautiful soup và đưa vào trong các list, từ list chuyển đổi thành dataframe.

Graphical user interface, text, application, email

Description automatically generated

Tuy nhiên, dữ liệu lấy được từ API Google Places lại là dạng JSON, lists of Dictionary , cần ETL rất nhiều, nhưng hiệu suất Query lại rất thấp do Google chỉ cho query tối đa 20 kết quả 1 lần, 80 lần là tối đa cho 1 địa điểm. Nên nếu cố gắng query nhiều lần để lấy nhiều kết quả có thể dẫn đến bị duplicate hoặc bị miss dữ liệu. Vì vậy nên em không sử dụng Google Places API mà sử dụng Foursquare APIGraphical user interface, text, application, email

Description automatically generated

Thành quả sau khi đã xử lý dữ liệu

Table

Description automatically generated

Dữ liệu hoàn chỉnh từ Foursquare API

A picture containing graphical user interface

Description automatically generated

Ngoài ra em cũng sử dụng Folium để marker và plot các quận huyện của TPHCM lên bản đồ, cùng với các cluster market sau đó Map

Description automatically generated

# Exploratory Data Analysis

Sau khi tiến hành một vài phương pháp phân tích dữ liệu cơ bản và thống kê , có những thông tin sau.

Những quận có nhiều điểm kinh doanh nhất

Chart, histogram

Description automatically generated

Những Quận có nhiều loại hình kinh doanh nhất

Chart

Description automatically generated

Những loại hình kinh doanh được gặp nhiều nhất

Table

Description automatically generated

Có thể thấy nhiều nhất chính là nhà hàng Việt Nam và quán ăn có kết hợp với uống cà phê ( như kiểu cơm trưa văn phòng ,... ) , thứ 3 là quán Cà Phê và thứ 4 là Khách sạn.

Lưu ý : vì khái niệm Café và Restaurant đôi khi nó giống nhau, nên ở Việt Nam có khi những nhà hàng Việt Nam cũng bị nhầm lẫn là Café. Tuy nhiên, đồ án này chỉ tập trung vào “ Vietnamese Restaurant” để có kết quả khách quan nhất.

Top 10 loại hình dịch vụ theo từng Quận

Table

Description automatically generated

Loại hình dịch vụ phổ biến nhất của các quận phần lớn vẫn là Nhà Hàng, quán ăn.

# Preprocessing

Trong phần đồ án này , task preprocessing chính vẫn là convert categorical data thành categorical interger data ( để phục vụ cho Kmean clustering) và convert ordinal data thành categorical data ( để phục vụ cho Kmode clustering)

Về task convert categorical data thành interget categorical data, sử dụng onehot encoder để biến những giá trị categorical như “Vietnamese Restaurant” ... thành những giá trị như 0 – tức no , 1 – tức yes. Sau đó sẽ dùng giá trị đó để xây dựng các cluster

Table

Description automatically generated with medium confidence

Về task convert ordinal data thành categorical data, sử dụng phương pháp IQR, tính phân bổ hay độ trải để chấm điểm và xếp hạng giá trị thành các categorical data.

Table

Description automatically generated

Các bước thực hiện :

1. Count số lần xuất hiện của venue

2. Tiến hành thống kê, xem xét phân bổ

3. Tiến hành chấm điểm mức độ

Kết hợp với sử dụng boxplot để convert

Chart, box and whisker chart

Description automatically generated

Graphical user interface, application

Description automatically generated

# Modelling

## Elbow Test

* Vì sử dụng thuật toán Clustering nên phải tìm điểm Elbow
* Thuật Toán Kmeans và Kmode đều cho kết quả Elbow tại k=4
* Kmean

Graphical user interface, text, application

Description automatically generated

* Graphical user interface

  Description automatically generated with low confidenceKmode

## Clustering

Map

Description automatically generatedKmeans

Củ Chi và Cần Giờ là 2 outlier

có 3 cluster rõ rệt :

+ Cluster 0 : những quận phát triển , ở trung tâm

+ Cluster 1 : những quận ở rìa thành phố, giáp với sân bay

+ Cluster 2 là Củ Chi : outliner nên tạm bỏ qua.

+ Cluster 3 là những quận/ huyện ngoại thành

Kmode

Map

Description automatically generated

Giờ đây Thủ Đức là 1 cluster riêng

Cần Giờ và Củ Chi không còn là outliner

3 cluster còn lại khác biệt rõ rệt :

+ Cluster 0 : những quận phát triển , ở trung tâm

+ Cluster 1 : những quận ở rìa trung tâm, ngoài ra có những huyện ở rìa thành phố nhưng có mật độ dân số cao cũng được vào cluster 1

+ Cluster 2 là Thủ Đức: Thủ Đức đứng 1 cluster riêng cũng hợp lý vì đây là quận mới lên Thành phố. Tương lai sẽ là một khu riêng biệt phát triển độc lập

+ Cluster 3 là những quận/ huyện ngoại thành

## Further Analysis

* Những yếu tố khác :
  + Average Housing Price ( chi phí bds)
  + Average Population Density (Mdds)
  + Total Companies (Tổng số công ty, văn phòng cơ quan,..)

Table

Description automatically generated with low confidence

# Findings

## 6.2 Discussion

Thực tế, ta chỉ muốn những quận có chi phí bất động sản vừa phải , hoặc "hời hơn" nếu so với mật độ dân số ( nghĩa là APD level cao hơn APH Level ), chứ không ai muốn những quận có chi phí bất động sản quá cao nhưng mật độ dân số thì lại thấp. Hoặc những quận có mật độ dân số cao , chi phí bất động sản cũng cao không kém thì cũng chẳng có lợi ích gì.

Sau khi phân tích các yếu tố phụ và mức độ cạnh tranh kết hợp với phương pháp Thống kê cơ bản. Em lọc được 3 quận có mối quan hệ giữa chi phí nhà đất và mật độ dân số khá hấp dẫn.

Text

Description automatically generated with medium confidence

AHP Level Low - APD Level High : Quận 12 ( chi phí nhà thấp, mật độ dân số cao )  
AHP Level Medium - APD Level High : Quận 8 ( chi phí nhà tương đối, mật độ dân số cao )  
AHP Level Medium - APD Level Very High : Quận 4 ( chi phí nhà tương đối , mật độ dân số rất cao)

A picture containing diagram

Description automatically generatedA picture containing diagram

Description automatically generatedA picture containing text

Description automatically generated

Nhắc lại ý nghĩa cluster :

\* Cluster 0 : Có ít cạnh tranh.

\* Cluster 1 : Có sự cạnh tranh tương đối.

\* Cluster 2 : Có sự cạnh tranh cao

Mặc dù quận 8 có giá bất động sản ở mức tương đối và mật độ dân cư cao, tuy nhiên vẫn không hấp dẫn do không có nhiều công ty và nhà xưởng hơn quận 4 là mấy, và mức cạnh tranh cũng rất cao

**Việc mở quán ăn ở quận 4, mặc dù có lợi thế là giá bất động sản ở mức rẻ hơn so với quận 8 và mật độ dân cư cao hơn, tuy nhiên, mật độ công ty / nhà xưởng trong khu vực này cũng không nổi trội hơn quận 8 và mức cạnh tranh cũng rất cao không khác biệt nên khá rủi ro khi kinh doanh tại đây.**

**việc mở 1 quán ăn Việt Nam ở quận 12 sẽ là tốt nhất vì có mật độ dân số tương đối, mật độ công ty , nhà xưởng vào mức cao trong khi giá bất động sản lại chỉ bằng 3/5 so với quận 4 và quận 8**

**Vậy , lựa chọn tốt nhất là nên mở một quán ăn Việt Nam ở quận 12 vì chi phí bất động sản rẻ, mật đô dân cư, công ty & nhà máy xí nghiệp tương đối cao và mức độ canh tranh thấp.**

## 6.2 Conclusion

Mục đích của Đồ Án này là tìm ra quận hợp lý nhất để mở một nhà hàng Việt Nam dựa trên các tiêu chí : Chi Phí cố định vừa phải & kinh tế ( chi phí bất động sản ), mật độ dân cư đông và số lượng công ty , nhà máy xí nghiệp nhiều để không chỉ kinh doanh cho dân địa phương mà còn cho công nhân của các nhà máy , nhân viên của các công ty do các đối tượng này có xu hướng ra ngoài ăn trưa, ghé ăn sáng ,... hoặc đi làm về ghé ăn nhậu. Sau khi tiến hành phân tích dữ liệu từ nhiều nguồn và xây dựng mô hình phân tích cụm thì quận hợp lý nhất để mở một nhà hàng Việt Nam là quận 12.

## 6.3 Comments

Ngoài ra, Kết quả này là một sự dự đoán khách quan từ dữ liệu kết hợp với cả suy luận theo tình hình thực tế dựa trên kết quả của dữ liệu. Đương nhiên sẽ không thể tránh khỏi những thiếu sót như mật độ quán ăn chưa phản ánh đúng thực tế ( vì tình hình quán ăn của Việt Nam đôi khi không có trong dữ liệu của Foursquare, hoặc dữ liệu từ Foursquare dán nhãn sai, ví dụ như quán ăn cơm trưa văn phòng Việt Nam nhưng lại dán nhãn Cà phê,... Tuy nhiên ban đầu việc collect dữ liệu đã lấy ngẫu nhiên 100 địa điểm ở từng quận, nếu quận nào có mật độ nhà hàng Việt Nam nhiều đương nhiên kết quả sẽ ra nhiều nên cũng hạn chế được vấn đề này.