MỞ ĐẦU

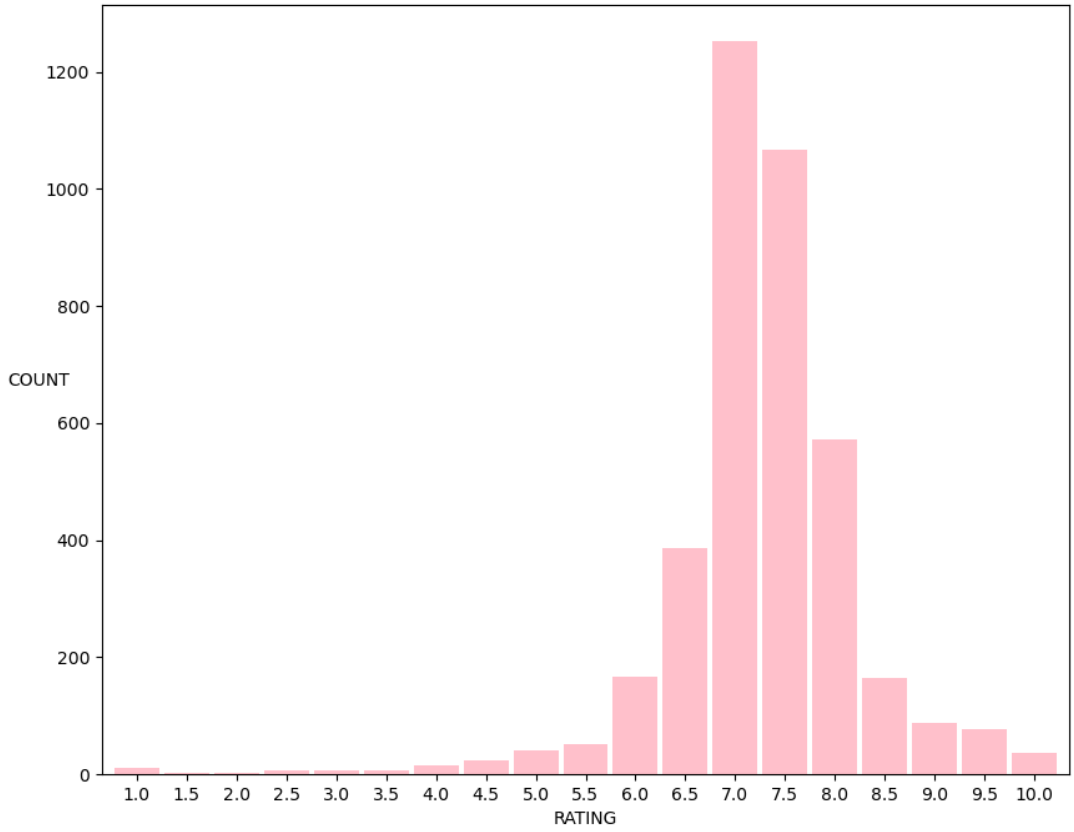
Bộ dữ liệu của nhóm gồm nhiều field, cho thấy đây là dự liệu có dimension lớn, cần phải cẩn thận phân tích để có thể trình chiếu trên 2D sao cho mang lại insight hữu ích, bên cạnh đó còn phải biết sử dụng tối đa dữ liệu quý cũng như biết nhận diện và loại bỏ dữ liệu nhiễu, dự liều thừa, dữ liệu hỏng.

Khổ nỗi là nhóm chưa có nhiều kinh nghiệm, clean xong vẫn chưa thấy dữ liệu nào đáng bỏ, ngoài ra cũng chưa rà soát nhiều lần để bắt các sample lỗi, cơ mà thôi deadline nên cứ dùng, hehe

PHÂN TÍCH

Phần 1 : Làm quen bộ dữ liệu

A. Phân bố rating tổng thể



Bước đầu ta xét một biểu đồ cột đơn giản, thể hiện số lượng của từng rating

Vì là bước đầu, ta đành gom rating của tất cả các quận, các món, các khoảng giá vào tính chung, tuy hơi kì nhưng từng bước sẽ dần khai phá, cốt để xây dựng cái sense về bộ dữ liệu.

Vì dữ liệu rating trải từ 0 tới 10 và chia nhỏ tới 0.1, nên để dễ hình dung, ta làm tròn rating tới đơn vị nhỏ nhất 0.5

Vì làm tròn tới 0.5 nên 0.7 sẽ được làm tròn xuống 0.5 (vì gần 0.5 hơn, hơi khó chịu tí sorry, nếu ta thích làm tròn 0.5 lên thì có thể chỉnh source code được cơ mà nó sẽ hơi bias=]]) còn 0.8 sẽ được làm tròn lên số tự nhiên tiếp theo (vì gần số mới hơn)

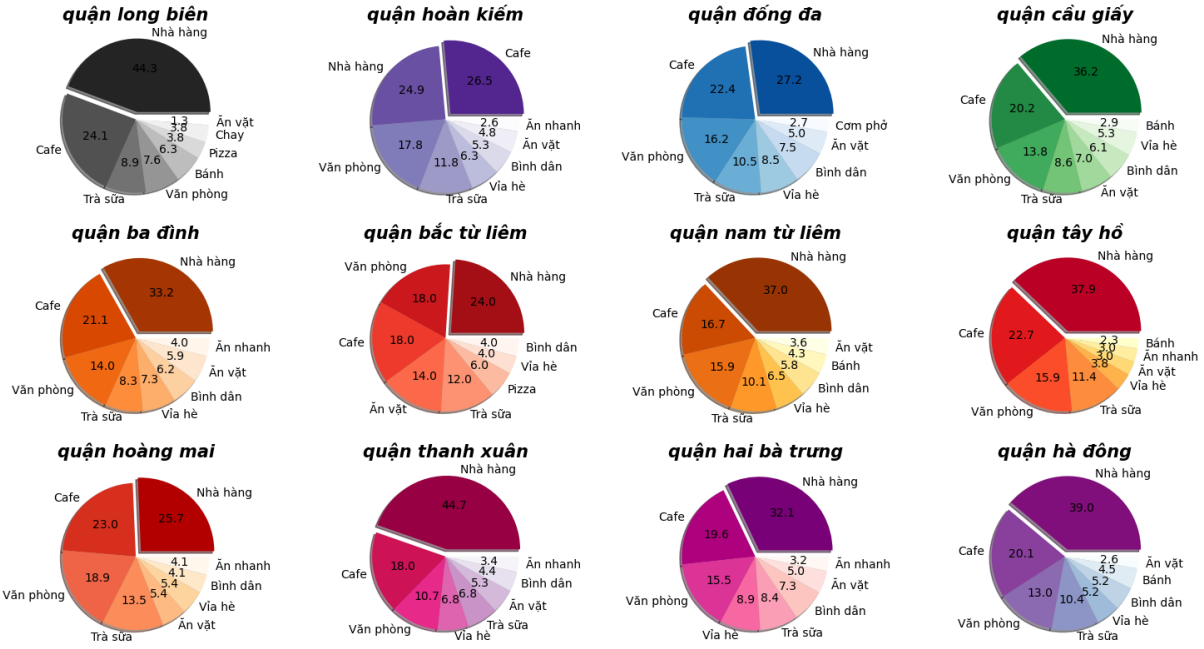
Vì để con người có được cái sense một cách dễ nhất về bộ dữ liệu, nên sẽ làm tròn như trên và dùng biểu đồ cột, thay vì dùng histogram và chọn size bin cho histogram. Histogram tuy sẽ đúng về mặt toán học thống kê hơn, nhưng vì mục tiêu của đề tài này không quá khắt khe, nên lựa chọn cách làm trên là hợp lí vì histogram khi plot lên trông hơi ngu ngu.

Tuy biểu đồ trên là tổng hợp rating của tất cả các món ăn, nên vẫn còn khá trừu tượng. Tuy vậy, bước đầu qua đó ta có thể nhận thấy rằng phân bố rating từ 7 cho tới 7.7 (dễ nhầm hơn histogram chưa, hehe) là rõ nhất, không chỉ vậy, nó vô cùng lớn so với các rating còn lại. Điều này khá hợp lí theo trực giác, nhưng qua đó ta cũng nhận thấy rằng để cung cấp và duy trì được dịch vụ đạt 9-10 điểm rating có thể nói khá là khó khăn.

Vì rating 7.5 theo trực giác được coi là “tương đối tốt”, mà dựa theo biểu đồ trên, 7.5 cũng là rating được xuất hiện tại trung tâm của curve, ta có thể tự tin assume rằng : khi nghiêng về phía bên phải của 7.5, sẽ là những đánh giá cho dịch vụ “chất lượng”.

B. Lấy ra top món ăn rating cao

Dựa vào assumption đó, ta lập biểu đồ thống kê ra TOP 8 loại hình dịch vụ (món ăn) ở từng quận, mà trong đó là các món có “số lượng rating lớn hơn hoặc bằng 7.5” cao nhất ở riêng quận đó.

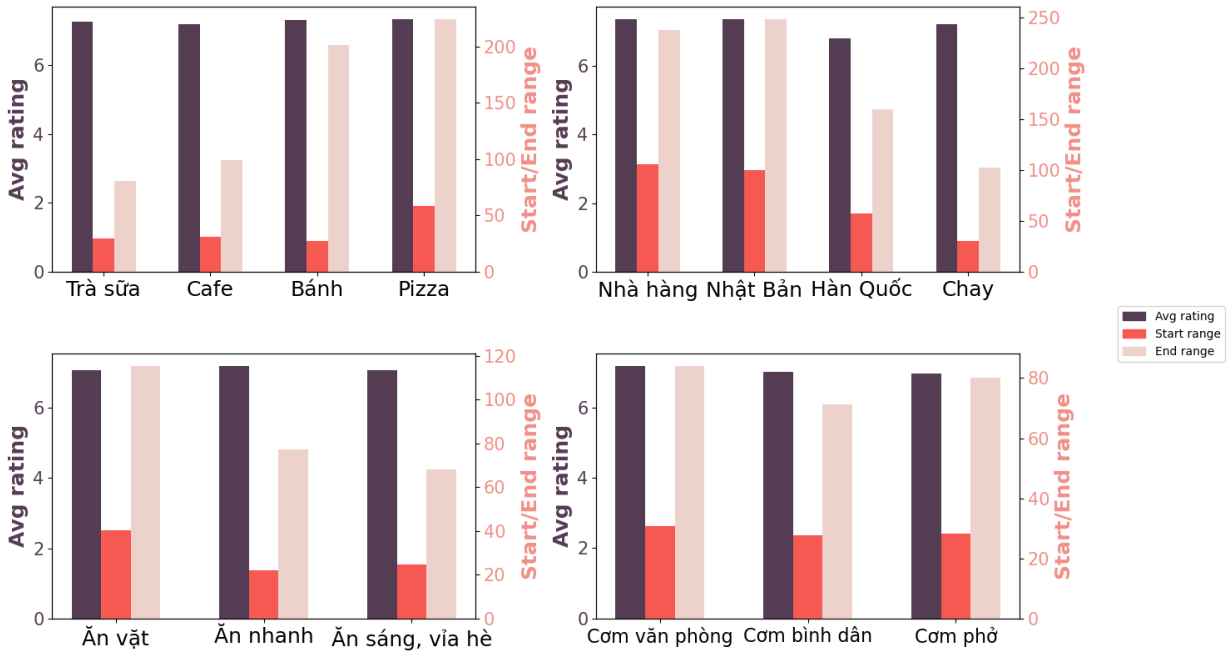


Qua biểu đồ quạt trên, ta thấy rằng phần lớn loại hình "chất lượng" là : nhà hàng, cafe, và cơm văn phòng - Khác với suy nghĩ rằng "các món ăn dễ có rating cao là nhờ độ đa dạng phong phú như : đồ hàn, đồ nhật, đồ ăn vặt, v.v " của đại đa số chúng ta.

Phần 2 : Định hình thô về bộ dữ liệu

A. Gom nhóm các món ăn

Nếu chỉ dựa vào số lượng rating cao thì chưa đủ để đưa ra các quyết định cần thiết, ta cần có thêm góc nhìn về khoảng giá, sau đây ta có multiple bar chart :



Thoạt nhìn qua, ta có thể thấy có 4 nhóm đồ ăn chính, đó là : nhà hàng, cơm phở, đồ ăn nhanh, đồ ăn……..không nhanh nhưng ít ăn thường xuyên????

4 nhóm dường như có chung tầm giá sàn nhưng khác nhau rất nhiều về giá trần (đây có lẽ là đặc điểm để phân loại 4 nhóm), ngoài ra, đúng như dự đoán ở biểu đồ trừu tượng đầu bài, ở đây ta có thể thấy rating trung bình cũng là vào khoảng 7 tới 7.5

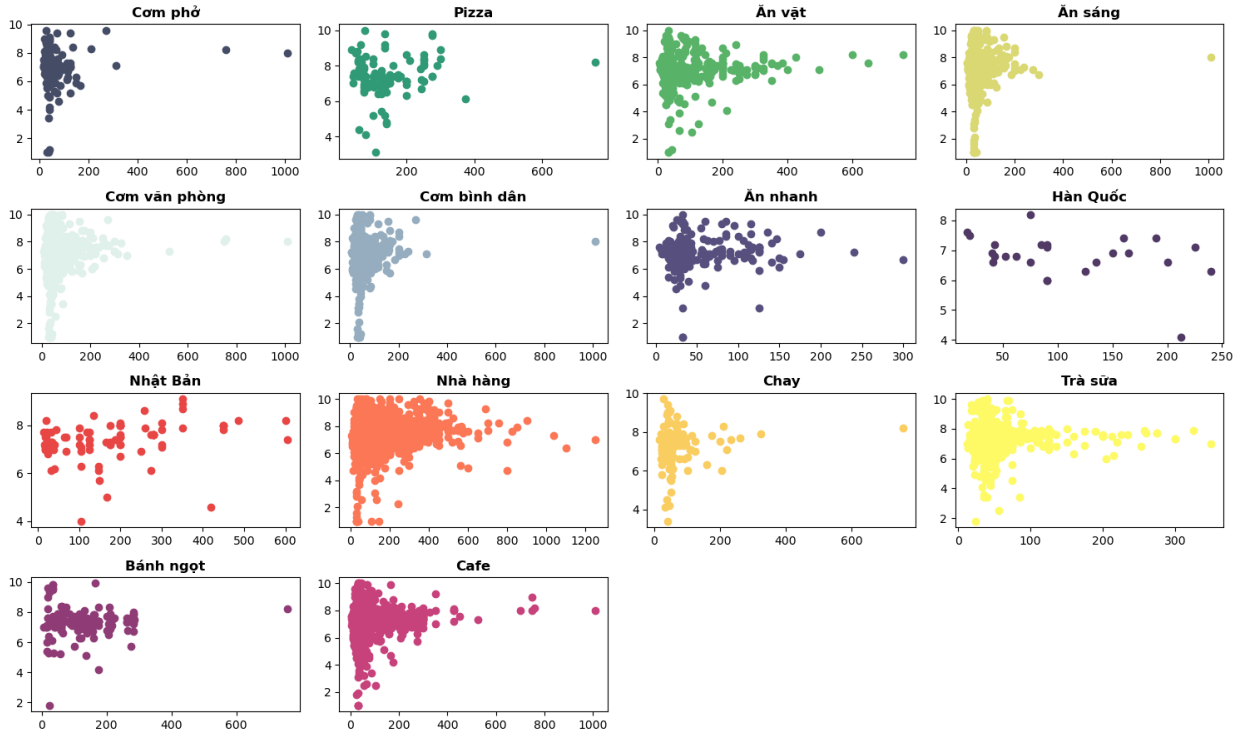
**** rating trung bình ở biểu đồ này là trung bình tất cả rating của món đó, sử dụng hàm mean() ***

B. Phân bố thô "rating và giá"

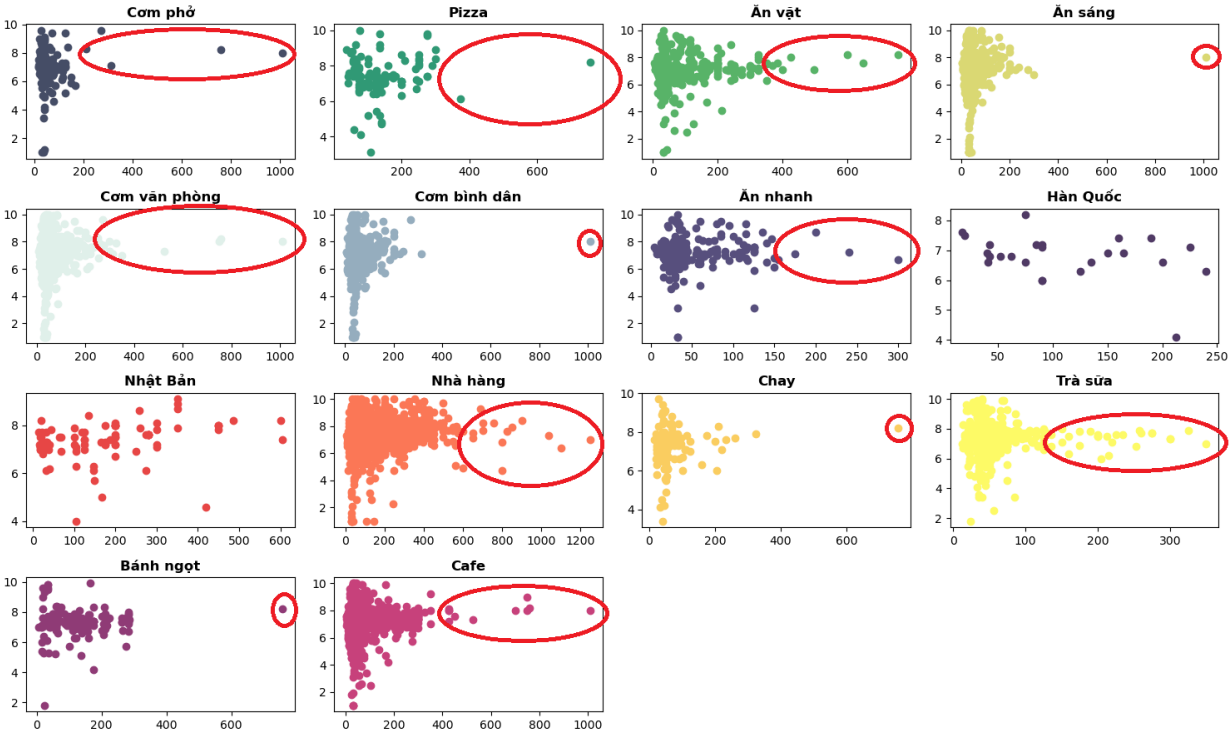
Dưới đây là biểu đồ scatter, vì là biểu đồ scatter nên sẽ sử dụng rating nguyên bản của bộ dữ liệu.

Nhưng vì bộ dữ liệu sử dụng giá tiền ở 2 đầu mút, không có 1 giá cố định nên không thể plot. Thay vào đó, ta sẽ lấy trung bình cộng "giá trần và giá sàn" của từng nơi bán để plot. Vì mỗi *biểu đồ con* là biểu diễn của riêng một món ăn, vì vậy từng cửa hàng sẽ sử dụng giá trung bình của riêng cửa hàng đó, vì nó giúp ta có cái nhìn về sự phân bố kĩ nhất, đủ nhất, thay vì phải lo tính tần suất của mỗi giá rồi mới tính trung bình v.v

Theo biểu đồ multiple bar bên trên, vì khoảng chênh lệch giữa giá trần và giá sàn tương đối lớn (với người nghèo như mình), nên giá trung bình chưa chắc phản ánh giá thật (món hàng bán chạy ngoài đời thực), nhưng vì làm gì có dự liệu nào khác để dựa vào nữa ôi bực thế nhờ, nên thôi cứ dùng như vậy, vì như đã nói ở trên, mỗi *biểu đồ con* chỉ biểu diễn phân bố của *một loại món ăn* nên chắc sẽ vẫn khách quan ít nhiều nào đó.



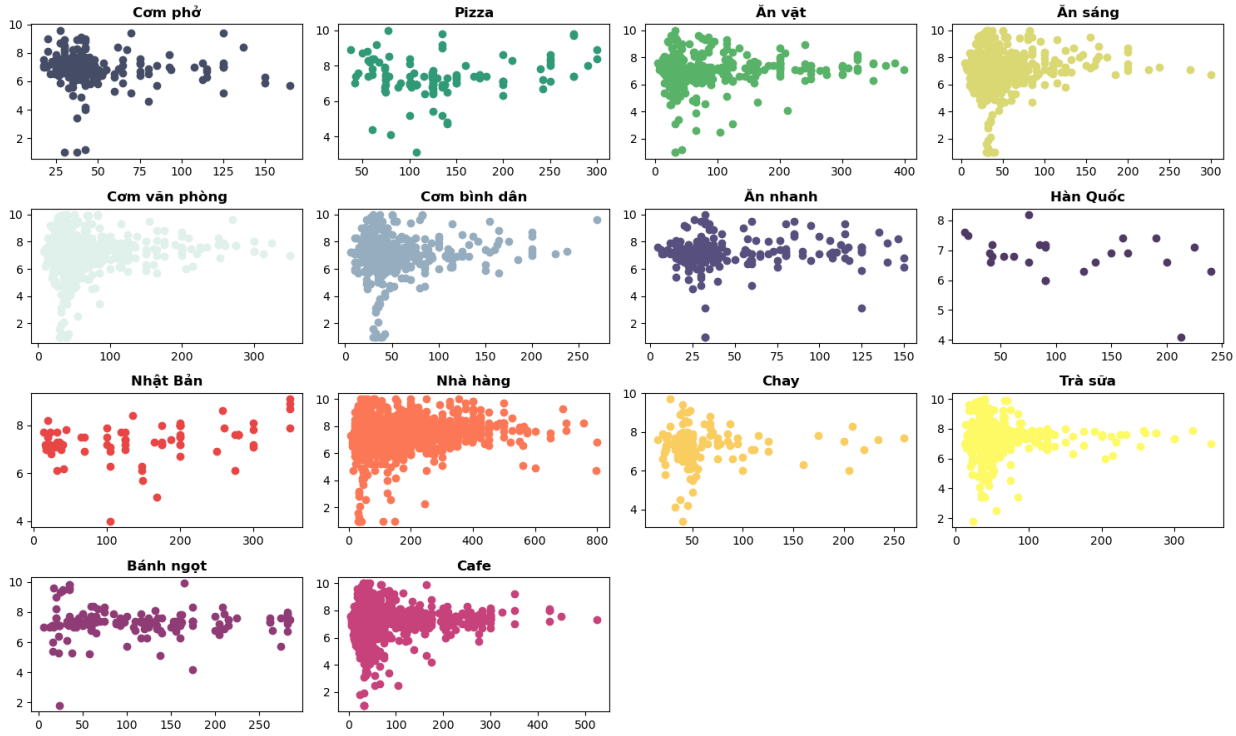
Qua biểu đồ này ta thấy vẫn còn một vài outliers, tần xuất xuất hiện của chúng ít nên ta sẽ tạm thời bỏ qua. Mặc dù ta nên giữ để xem xét nếu các outliers thuộc loại "chất lượng" – tức điểm rating 8 đến 10, nhưng vì số lượng quá ít – tức ít quán bán giá đó, và do ta đánh giá bằng mắt chứ không cho vào mô hình học máy để tính toán siêu kĩ lưỡng (vì đề tài đơn giản) nên ta sẽ chú tâm vào 2 khía cạnh : nơi tập trung đông + rating "chất lượng". Vì ta "cần" chạy theo số đông.


Cơm phở
Pizza
Ăn vặt
Ăn sáng
Cơm văn phòng
Cơm bình dân
Ăn nhanh
Hàn Quốc
Nhật Bản
Nhà hàng
Chay
Trà sữa
Bánh ngọt
Cafe

Phần 3 : Lọc kĩ và đưa ra nhận xét

A. Chọn ra tổ hợp phù hợp

Sau khi lọc bỏ, ta được :



Ở biểu đồ này, ta rút ra một vài đánh giá như sau:

- quả thực rating đúng là tập trung ở 7 và 7.5

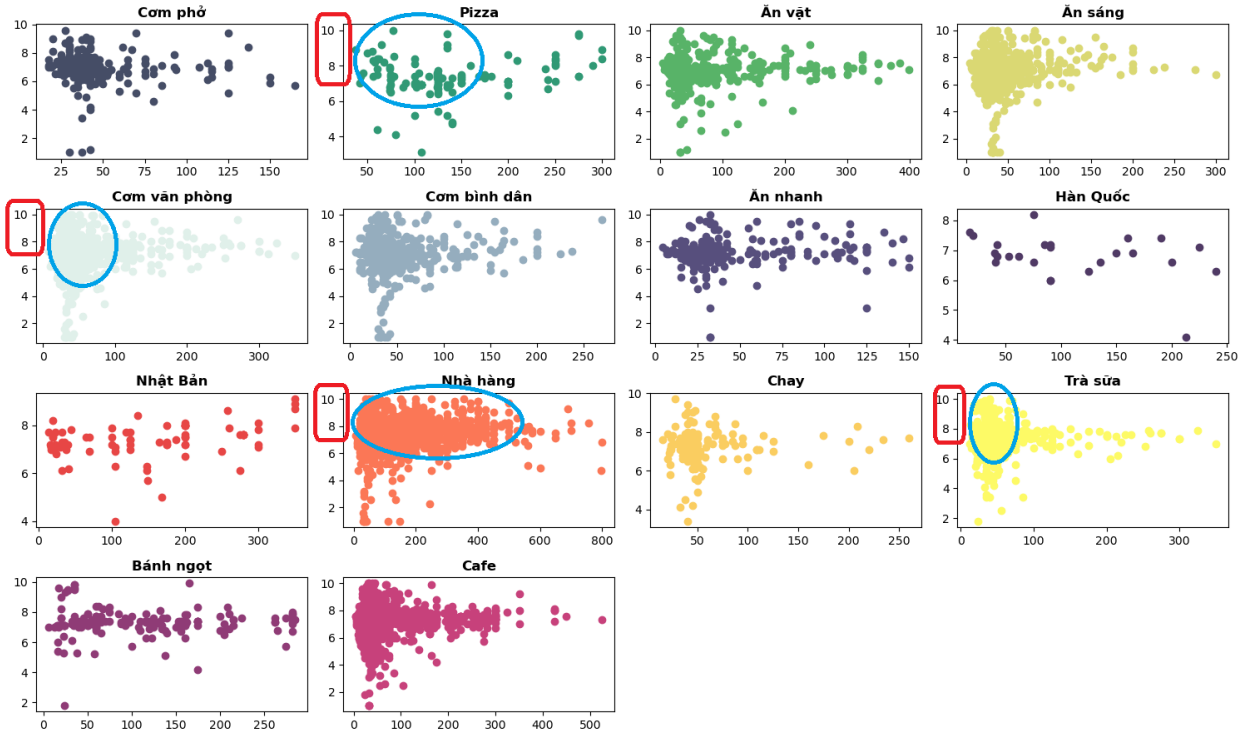
- giá thành bình dân (dưới 150k) tập trung nhiều nhất, xuất hiện ở mọi rating

- giá cao ít bán (→ ít người mua) (mà đã loại bớt outliers rồi nhé)

- giá cao tập chung quanh ở rating "tạm ổn" – tức 7 tới 7.5 → qua đó cho thấy giá cao cũng không phải quá thất vọng (không xuất hiện nhiều rating thấp), nhưng cũng chưa phải đồng biến đúng theo giá tiền (tiền nào của nấy - expecting giá cao là rating cao) → tạm chấp nhận(với người mua) và nên cân nhắc (với người bán), vì còn các vấn đề khác như : lời lãi, vệ sinh an toàn,v.v

====> để an toàn, ta nên lựa chọn "khoảng giá bình dân" và "rating tốt"

Ta lấy ví dụ thử nghiệm ở 4 biểu đồ con :



-Pizza : nên bán(mua) ở khoảng giá 50k-150k

-Cơm văn phòng : nên bán(mua) ở dưới 100k

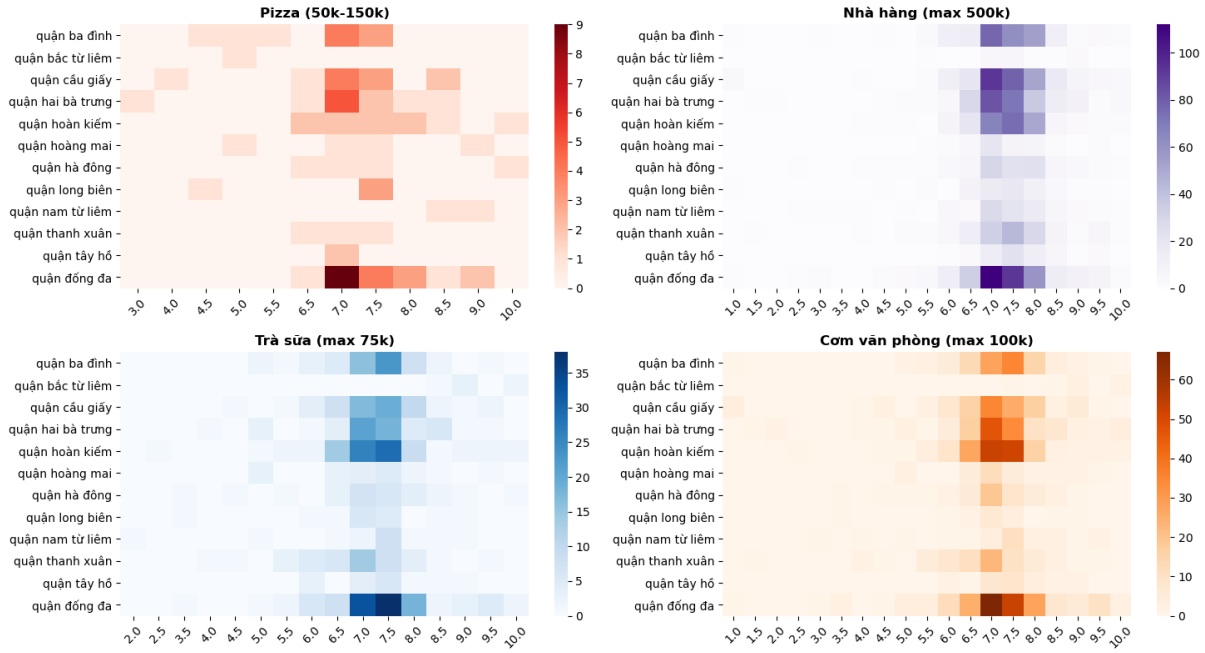
-Nhà hàng:nên bán(mua) ở dưới 500k

-Trà sữa : nên bán(mua) ở dưới 75k-80k

*** “nên bán” ở đây tức là lựa giá món hoặc tổng giá nhiều món bán chạy sao cho ở trong khoảng giá đó***

B. Đưa ra những quyết định cuối cùng về các vấn đề liên quan tới rating

Dựa vào 4 ví dụ đó, ta tạo heatmap:

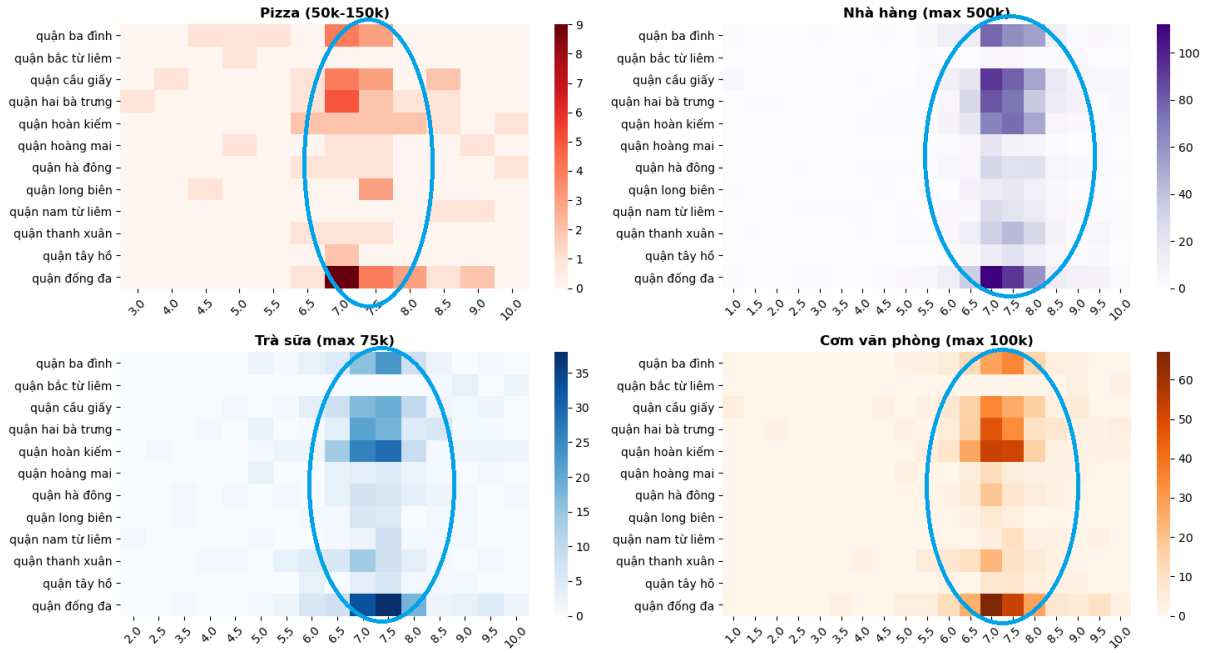


Heatmap ở đây sử dụng rating làm tròn 0.5 (như biểu đồ đầu tiên) thay vì rating nguyên bản như biểu đồ scatter phía trên (vẫn vì lí do là để giúp con người DỄ HIỂUUUU)

Ở đây màu đậm thể hiện số lượng càng nhiều rating-ứng-với-trục-hoành

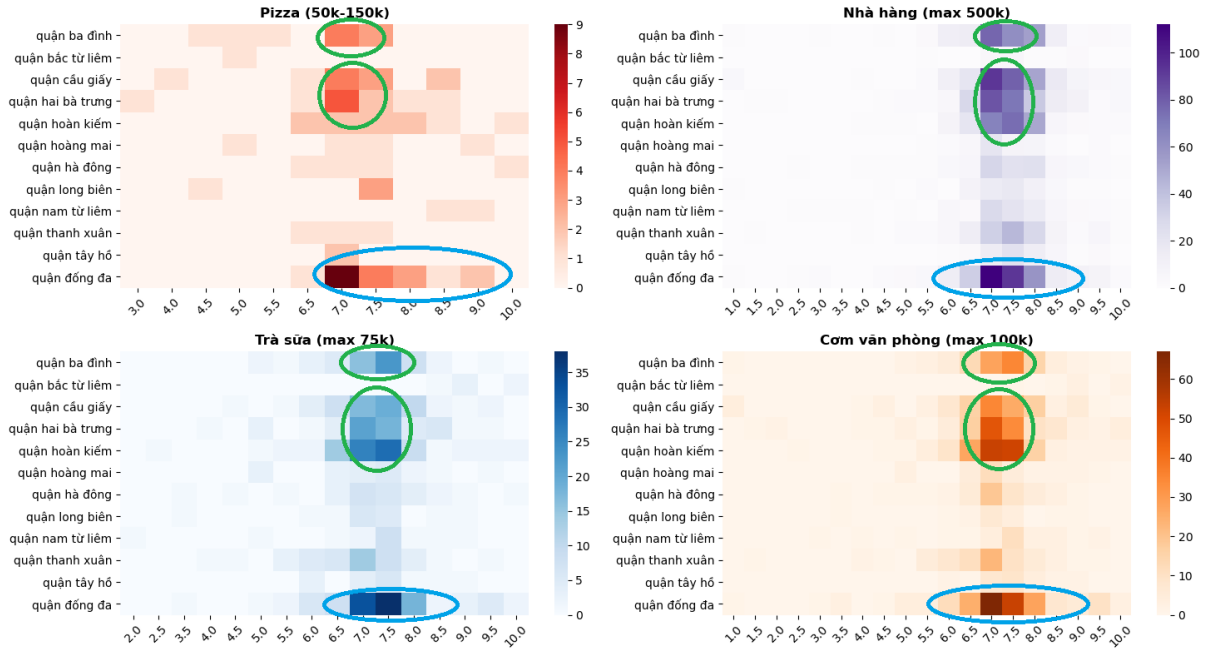
Trước khi tạo heatmap, ta phải lọc đúng khoảng giá tương ứng với vùng "cần dùng" như đã khoanh ở biểu đồ scatter phía trên. Ví dụ như: pizza chỉ lấy sample mà giá chạy từ 50k tới 150k, nhà hàng thì chỉ lấy sample dưới 500k, v.v (giá trung bình cộng của trần và sàn nhé)

Nhìn vào heatmap, ta có nhận định đầu tiên :



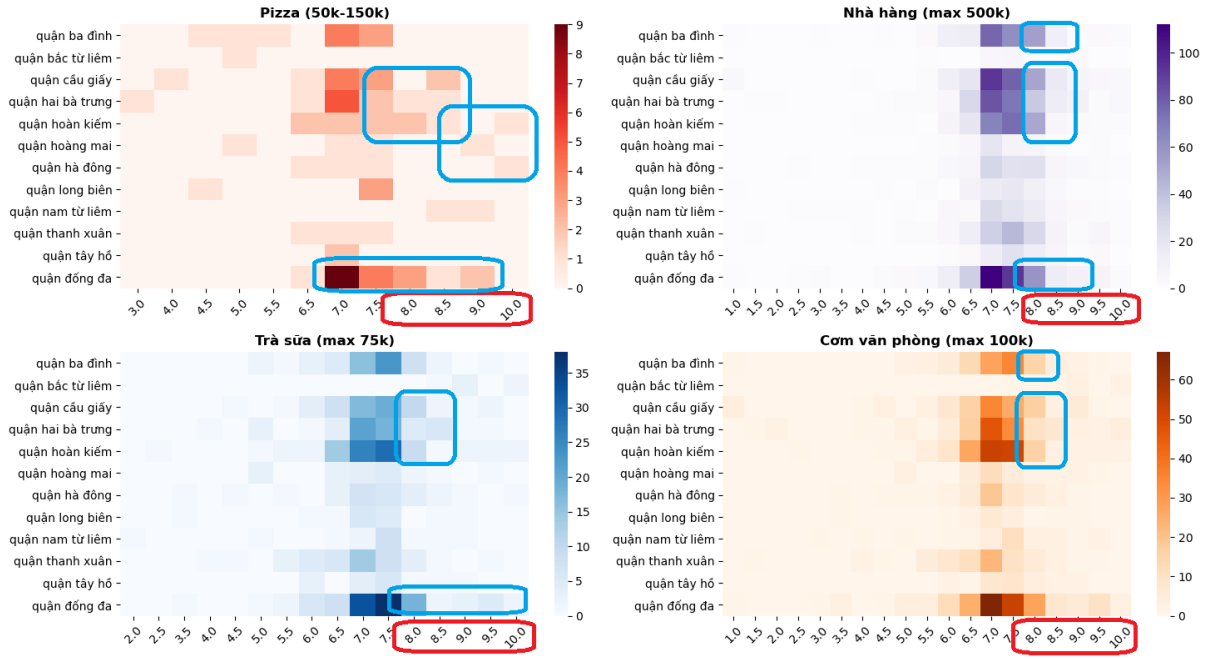
-một lần nữa, đúng là khoảng giá 7 – 7.5 là xuất hiện nhiều nhất, và vô cùng vượt trội, ở mọi món (ở đây ta lấy ra 4 món làm ví dụ)

Ta có nhận xét thứ 2 :



-những quận có được NHIỀU (đậm) rating từ “tạm ổn” tới “tốt” tập trung ở các quận : Ba Đình, Cầu Giấy, Hai Bà Trưng, Hoàn Kiếm, và tiêu biểu nhất có thể thấy rõ là Đống Đa (hơi điêu điêu…….chắc do data sai, xin lỗi các tình yêu=]])

Quay trở lại vấn đề, sau khi đã "lọc" khoảng giá cần thiết, ta còn 1 bước nữa là phải chọn "rating tốt" như đã khoanh ở scatter plot. Ta sẽ chọn ở heatmap này :



Ta thấy, tiêu biểu vẫn là các quận vừa nêu ở phía trên, và đặc biệt nếu phải chọn thì cả người mua lẫn người bán nên cân nhắc Quận Đống Đa cho 4 loại món trên.

TẠM KẾT

Từ phần analysis trên, ta có thể cảm nhận được phần nào về bộ dự liệu một cách rõ ràng, có cái nhìn tốt hơn khi đưa ra quyết định cho cả người mua và người bán, ta hiểu được một số bước cần thực hiện để tìm tới mối tương quan giúp cho việc so sánh được tốt hơn. Qua đó, ta thấy ta đã tiến thêm được kha khá so với dữ liệu thô đầy trừu tượng ban đầu. Tuy vậy, bộ dự liệu này vẫn còn nhỏ, còn chưa khách quan, nhân lực lại chưa có nhiều kinh nghiệm và kĩ năng phân tích dữ liệu, cần phải học hỏi và cải tiến.