**TRƯỜNG ĐẠI HỌC GIAO THÔNG VẬN TẢI TP. HỒ CHÍ MINH**

**KHOA CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO THỰC TẬP TỐT NGHIỆP**

**Tên đề tài:**

**Phân tích dữ liệu bán hàng, đánh giá hiệu quả trong quá trình kinh doanh**

**Giảng viên hướng dẫn**: ThS. Phan Thị Hồng Nhung

**Sinh viên thực hiện:**

*20H1120258: Bùi Đức Thịnh*

*20H1120175: Đinh Hoàng Thịnh*

Mục lục

[**Lời mở đầu** 1](#_Toc150853532)

[**Danh mục từ viết tắt** 1](#_Toc150853533)

[**I. Tổng quan:** 1](#_Toc150853534)

[**1.1. Ý tưởng thực hiện đề tài:** 1](#_Toc150853535)

[**1.2. Mục tiêu thực hiện đề tài:** 2](#_Toc150853536)

[**1.3. Phương pháp nghiên cứu:** 2](#_Toc150853537)

[**1.4. Kết quả đạt được khi hoàn thành đề tài:** 2](#_Toc150853538)

[**II. Nội dung chính:** 2](#_Toc150853539)

[**2.1. Giới thiệu chung về một số công cụ và ngôn ngữ sử dụng để phân tích:** 2](#_Toc150853540)

[**2.1.1. Python:** 2](#_Toc150853541)

[**2.1.2. SQL:** 3](#_Toc150853542)

[**2.1.3. PowerBI:** 5](#_Toc150853543)

[**2.1.4. Excel:** 5](#_Toc150853544)

[**2.2. Thực hiện triển khai đề tài:** 6](#_Toc150853545)

[**2.2.1. Làm sạch loại bỏ dữ liệu dư thừa:** 6](#_Toc150853546)

[**2.2.2. Sử dụng SQL thực hiện truy vấn dữ liệu theo yêu cầu:** 14](#_Toc150853547)

[**2.2.4. Mô hình hóa dữ liệu:** 17](#_Toc150853548)

[**2.2.5. Đưa ra những nhận định, phân tích tổng quát dữ liệu:** 18](#_Toc150853549)

[**III. Kết quả dự án:** 21](#_Toc150853550)

[**3.1. Đánh giá chung:** 21](#_Toc150853551)

[**3.2. Những thiếu sót và hạn chế trong quá trình thực hiện đề tài:** 21](#_Toc150853552)

[**3.3. Đề xuất phát triển và cải thiện:** 21](#_Toc150853553)

[**IV. Kết luận:** 22](#_Toc150853554)

[**V. Tài liệu tham khảo:** 22](#_Toc150853555)

# **Lời mở đầu**

Với việc ứng dụng công nghệ thông tin vào tất cả lĩnh vực như hiện nay việc áp dụng công nghệ vào kinh doanh cũng không nằm ngoài cuộc, với phương pháp đánh giá kinh doanh truyền thống thì hiệu quả đem lại không cao nhưng khi ta áp dụng công nghệ vào việc đánh giá kinh doanh bán hàng đưa dữ liệu bán hàng từ những dữ liệu thô sơ ta có thể trực quan hóa dữ liệu giúp cho những người quản lý tổng quát được thông tin từ đó có thể đánh giá xem xét cải thiện tốt hơn việc đánh giá hiệu quả kinh doanh so với việc sử dụng dữ liệu bán hàng truyền thống, từ đó giúp cho việc kinh doanh ngày càng hiệu quả và phát triển hơn. Với mong muốn trở thành nhà phân tích dữ liệu (Data analyst) nhóm đã sử dụng dữ liệu có sẵn từ cửa hàng bán bánh pizza, qua việc làm sạch phân tích và trực quan hóa bộ dữ liệu, mục tiêu của nhóm là phân tích đánh giá bộ dữ liệu có sẵn này để có thể đưa ra những phương án tối ưu nhất giúp cho cửa hàng có thể phát triển cải thiện doanh thu cho cửa hàng. Vì chưa có kinh nghiệm nhiều trong việc phân tích nhận định và đánh giá về dữ liệu nhưng nhóm sẽ cố gắng hoàn thiện bài phân tích báo cáo trong khả năng của mình một cách tốt nhất. Nhóm em xin chân thành cảm ơn cô Nhung giáo viên hướng dẫn cho nhóm, hi vọng cô sẽ giúp đỡ hỗ trợ nhóm em hơn trong quá trình thực hiện đề tài này.

# **Danh mục từ viết tắt**

|  |  |
| --- | --- |
| **Viết tắt** | **Đầy đủ** |
| PNG hoặc SGV | Định dạng file ảnh |
| RDBMS | Relational Database Managerment System (hệ quản trị cơ sở dữ liệu) |

# **I. Tổng quan:**

## **1.1. Ý tưởng thực hiện đề tài:**

Hiện nay với sự phát triển không ngừng của công nghệ thông tin trong đời sống về mọi mặt đời sống, xã hội, y tế, con người. Đi kèm với sự phát triển đó sẽ tạo ra nguồn dữ liệu vô cùng lớn, vì vậy nhờ vào những dữ liệu có sẵn ta có thể dự đoán xu hướng, nhu cầu tiêu dùng, mua sắm, hay sinh hoạt của con người để từ đó nhằm đưa ra những kết quả chính xác giúp nâng cao hiệu quả đời sống cho con người tốt hơn, chính vì những điều đã nêu trên đã làm cho nhóm cảm thấy rất hứng thú với việc thu thập làm sạch và dự đoán kết quả dựa vào nguồn dữ liệu có sẵn và đó là lý do nhóm thực hiện đề tài.

## **1.2. Mục tiêu thực hiện đề tài:**

Vì muốn đánh giá nhu cầu tiêu dùng của khách hàng nên ở đây nhóm dùng dữ liệu của một cửa hàng bán bánh pizza, dựa vào dữ liệu có sẵn nhóm em sẽ thực hiện lọc, đánh giá sức mua, số lượng bánh bán ra, loại bánh nào có doanh thu cao nhất,… để từ đó nhằm nâng cấp cải thiện hơn về chất lượng sản phẩm đã bán.

## **1.3. Phương pháp nghiên cứu:**

Phương pháp nghiên đề cập đến một cách tiếp cận có hệ thống, tiến hành thu thập dữ liệu bằng Excel kết hợp làm sạch bằng ngôn ngữ lập trình Python và cuối cùng là trực quan hóa dữ liệu bằng công cụ PowerBI, từ đó phân tích kết quả và rút ra kết luận dựa trên bộ dữ liệu có sẵn. Bằng cách làm theo phương pháp này, nhóm có thể giảm thiểu sự dư thừa trong dữ liệu, tối ưu hóa chi phí bán hàng và sai sót trong quá trình bán hàng, giúp đảm bảo kết quả nghiên cứu là đáng tin cậy và hợp lệ.

## **1.4. Kết quả đạt được khi hoàn thành đề tài:**

Qua quá trình thu thập, đánh giá, phân tích và trực quan hóa dữ liệu kết quả sẽ giúp cho dữ liệu sạch, đánh giá được hiệu quả bán hàng, xác định được nhu cầu và sở thích tiêu dùng của khách hàng, cuối cùng là làm cho dữ liệu được trực quan để có thể dễ dàng đánh giá phát triển hơn trong tương lai.

# **II. Nội dung chính:**

## **2.1. Giới thiệu chung về một số công cụ và ngôn ngữ sử dụng để phân tích:**

### **2.1.1. Python:**

Python là một ngôn ngữ lập trình được sử dụng rộng rãi trong các ứng dụng web, phát triển phần mềm, khoa học dữ liệu và máy học (Machine Learning). Các nhà phát triển sử dụng Python vì nó hiệu quả, dễ học và có thể chạy trên nhiều nền tảng khác nhau. Ở đây nhóm sử dụng một số thư viện hỗ trợ để có thể sử dụng phân tích làm sạch và trực quan dữ liệu.[1]

#### **2.1.1.1. Thư viện hỗ trợ lập trình:**

* Pandas: Pandas là một thư viện Python cung cấp các cấu trúc dữ liệu nhanh, mạnh mẽ, linh hoạt và mang hàm ý. Tên thư viện được bắt nguồn từ panel data (bảng dữ liệu). Pandas được thiết kế để làm việc dễ dàng và trực quan với dữ liệu có cấu trúc (dạng bảng, đa chiều, có tiềm năng không đồng nhất) và dữ liệu chuỗi thời gian.

+ Pandas rất phù hợp với nhiều loại dữ liệu khác nhau:

Dữ liệu dạng bảng với các cột được nhập không đồng nhất, như trong bảng SQL hoặc bảng tính Excel.

+ Dữ liệu chuỗi thời gian theo thứ tự và không có thứ tự (không nhất thiết phải có tần số cố định).

+ Dữ liệu ma trận tùy ý (được nhập đồng nhất hoặc không đồng nhất) với nhãn hàng và cột.

+ Bất kỳ hình thức khác của các bộ dữ liệu quan sát / thống kê.

+ Dữ liệu thực sự không cần phải được dán nhãn vào cấu trúc dữ liệu pandas. Pandas được xây dựng dựa trên NumPy.

+ Hai cấu trúc dữ liệu chính của pandas là Series (1 chiều) và DataFrame (2 chiều) xử lý được phần lớn các trường hợp điển hình trong tài chính, thống kê, khoa học xã hội và nhiều lĩnh vực kỹ thuật.[2]

* Matplotlib: Là một thư viện vẽ đồ thị rất mạnh mẽ hữu ích cho những người làm việc với Python và NumPy. Module được sử dụng nhiều nhất của Matplotib là Pyplot cung cấp giao diện như MATLAB nhưng thay vào đó, nó sử dụng Python và nó là nguồn mở. Nó được dùng để trực quan hóa dữ liệu để có thể thực hiện các suy luận, thống kê.[3]

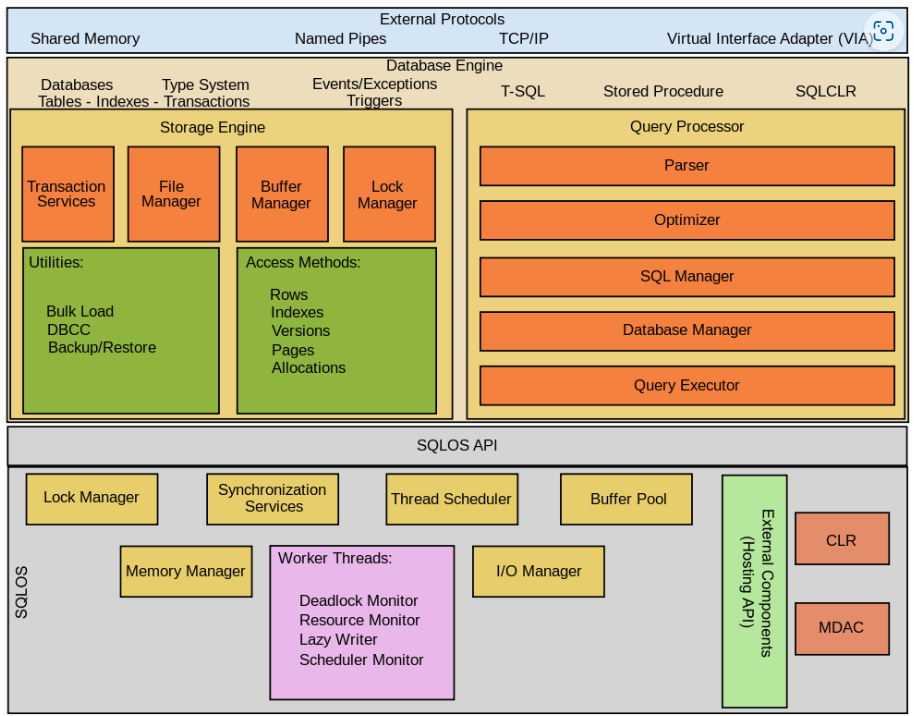
Không những thế thư viện này còn cho phép khai thác sử dụng kết hợp chung cùng với những thư viện khác có sẵn chẳng hạn như Numpy hay Pandas,… . Cuối cùng nó còn có thể biểu diễn xuất ra các dạng đồ họa phổ biến như PNG hoặc SGV để dùng cho các mục đích khác tùy vào nhu cầu người sử dụng.[4,tr.168]

### **2.1.2. SQL:**

SQL Server hay Microsoft SQL Server là phần mềm ứng dụng cho hệ thống quản trị cơ sở dữ liệu quan hệ (Relational Database Management System – RDBMS) được phát triển bởi Microsoft vào năm 1988. Nó được sử dụng để tạo, duy trì, quản lý và triển khai hệ thống RDBMS. Phần mềm SQL Server được sử dụng khá rộng rãi vì nó được tối ưu để có thể chạy trên môi trường cơ sở dữ liệu rất lớn lên đến Tera – Byte cùng lúc phục vụ cho hàng ngàn user. Bên cạnh đó, ứng dụng này cung cấp đa dạng kiểu lập trình SQL từ ANSI SQL (SQL truyền thống) đến SQL và cả T-SQL (Transaction-SQL) được sử dụng cho cơ sở dữ liệu quan hệ nâng cao.[5]

Cấu trúc của SQL Server bao gồm 2 thành phần: Database Engine và SQLOS

* Database Engine: Là thành phần chưa các cơ sở dữ liệu quan hệ, nó bao gồm Relation Engine giúp xử lý các truy vấn và Storage Engine giúp quản lý các tệp cơ sở dữ liệu, các trang, chỉ mục,…[5]
* SQLOS: Dưới Database Engine là hệ điều hành SQL Server hay SQLOS (SQL Server Operating System). SQLOS giúp quản lý bộ nhớ và I/O, lên lịch nhiệm vụ và khóa dữ liệu để tránh các xung đột xảy ra khi update,[5] chịu trách nhiệm quản lý tài nguyên hệ điều hành dành riêng cho SQL Server. [6]



**Hình 2.1:** Mô hình cấu trúc của SQL[6]

#### **2.1.2.1 Một số yêu cầu được đưa ra để phân tích:**

Trong quá trình phân tích ta xem xét, dựa vào dữ liệu có sẵn nhóm đã suy luận tự mình đưa ra những dự đoán trong quá trình phân tích từ đó có thể xem xét đánh giá khách quan nhất về bộ dữ liệu nhóm đang nghiên cứu. Một số dự đoán, yêu cầu mà nhóm đã đưa ra để phân tích bộ dữ liệu bán bánh pizza gồm có 2 phần là doanh thu và xu hướng mua hàng của khách được chia ra như sau:

1. **Doanh thu bán hàng:**

* Tổng doanh thu
* Tổng số đơn hàng đã bán được
* Giá trị trung bình cho mỗi đơn hàng
* Trung bình số lượng hàng bán ra qua mỗi đơn
* Tổng số bánh pizza có giá tiền tương ứng

1. **Xu hướng mua hàng của khách hàng**

* Tổng số đơn hàng đặt mỗi ngày trong tuần
* Tổng số đơn hàng đặt qua mỗi tháng
* Phần trăm số pizza bán ra theo loại (category)
* Phần trăm size pizza đã bán trong đơn hàng
* Top 5 pizza có doanh thu thấp nhất
* Top 5 pizza có doanh thu cao nhất
* Top 5 pizza có số lượng đơn bán ra nhiều nhất
* Top 5 pizza có số lượng đơn bán ra thấp nhất
* Số pizza bán được trong 1 ngày
* Số bánh bán ra trong 1 quý
* Số bánh bán ra trong 1 tháng
* Số size pizza bán ra trong tháng 1 đến tháng 12 bất kì

### **2.1.3. PowerBI:**

Power BI là công cụ phân tích và trực quan hóa dữ liệu tương tác của Microsoft dành cho lĩnh vực Business Intelligence (BI) (tạm dịch là phân tích dữ liệu thông tin doanh nghiệp). PowerBI là một dịch vụ phân tích kinh doanh của Microsoft nhằm mục đích cung cấp trực quan dữ liệu & thông tin thông minh kinh doanh (BI) với giao diện đủ đơn giản để người dùng cuối tạo báo cáo và trang tổng quan của riêng họ.[7]

Microsoft Power BI được sử dụng để tìm thông tin chi tiết trong dữ liệu của tổ chức. Power BI có thể kết nối nhiều loại tệp, chuyển đổi và trực quan hoá thành mô hình dữ liệu. Tạo biểu đồ, đồ thị để cung cấp hình ảnh trực quan và sự tương quan của dữ liệu. Tất cả những điều này có thể được chia sẻ với những người dùng Power BI khác trong tổ chức.[7]

Doanh nghiệp sử dụng mô hình dữ liệu từ Power BI để dự đoán. (“nếu có thể xảy ra/what if” trong dữ liệu). Các dự đoan theo thời gian thực để kịp thời phản ứng với các chỉ số kinh doanh.[7]

### **2.1.4. Excel:**

Microsoft Excel là chương trình xử lý bảng tính nằm trong bộ Microsoft Office của hãng phần mềm Microsoft được thiết kế để giúp ghi lại, trình bày các thông tin xử lý dưới dạng bảng, thực hiện tính toán và xây dựng các số liệu thống kê trực quan có trong bảng từ Excel.[8]

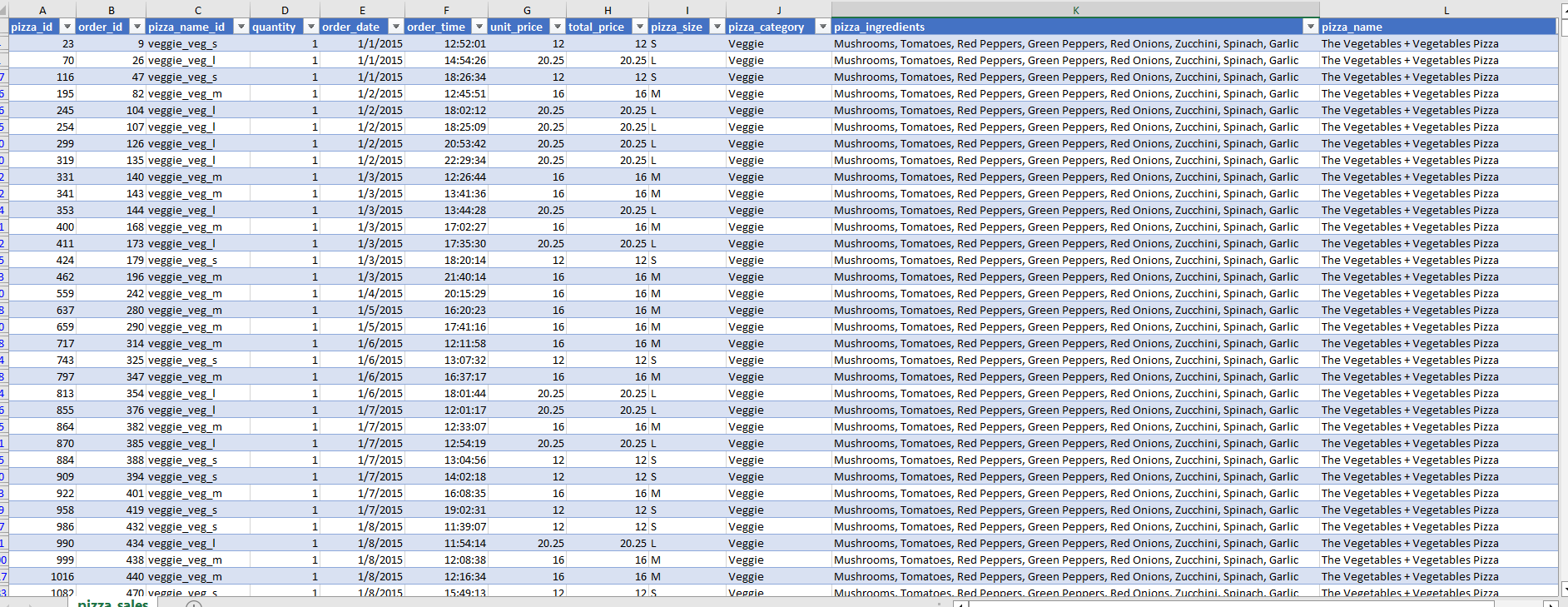
Sức mạnh của Excel còn được các lập trình viên tích hợp vào các công cụ phân tích và tìm kiếm rất nhiều. Đặc biệt có thể kể đến là các công cụ liên quan đến phân tích chẳng hạn như PowerBI hay ngôn ngữ lập trình Python. Với nguồn dữ liệu tăng lên liên tục theo thời gian cũng như khả năng lưu trữ ngày càng lớn thì việc lọc ra danh sách dựa theo các tiêu chí và xuất ra file Excel được rất nhiều người quan tâm.[8]

## **2.2. Thực hiện triển khai đề tài:**

### **2.2.1. Làm sạch loại bỏ dữ liệu dư thừa:**

#### **2.2.1.1. Thu thập dữ liệu:**

Đây là công đoạn đầu tiên trong quá trình thực hiện phân tích dữ liệu, đầu tiên nhóm đã thu thập dữ liệu bán bánh pizza dựa vào dữ liệu có sẵn trên trang web: [link](https://drive.google.com/drive/folders/17U0ah6Q4MJM_wIn_Xl4fHc-1fO6Q4s6z). Sau khi lấy dữ liệu trên về nhóm đã sử dụng Excel dùng để lưu trữ và xử lý dữ liệu ban đầu.



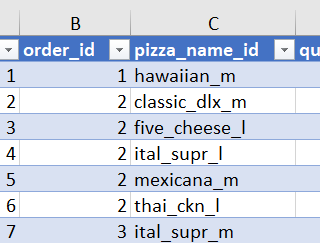
**Hình 2.2:** Dữ liệu ban đầu khi thu thập

Bộ dữ liệu trên là dữ liệu bán bánh pizza trong vòng 12 tháng chứa các cột bao gồm: pizza\_id (mã pizza), order\_id (mã đặt hàng), pizza\_name\_id (mã tên bánh), quantity (số lượng), order\_date (ngày đặt hàng), order\_time (thời gian đặt hàng), unit\_price (đơn giá), total\_price (tổng giá), pizza\_size (kích thước bánh), pizza\_category (loại bánh), pizza\_ingredients (nguyên liệu sử dụng trong bánh), pizza\_name (tên bánh). Tất cả các cột đều chứa thông tin mô tả thông tin của từng loại bánh cho đến đơn đặt hàng,… .

#### **2.2.1.2. Đánh giá tổng quát bộ dữ liệu:**

Sau khi xem xét đánh giá tổng quan qua bộ dữ liệu nhóm đã tìm và xác định được một số vùng dữ liệu bị dư thừa, đánh sai chữ hoặc trùng dữ liệu được mô tả như hình sau:

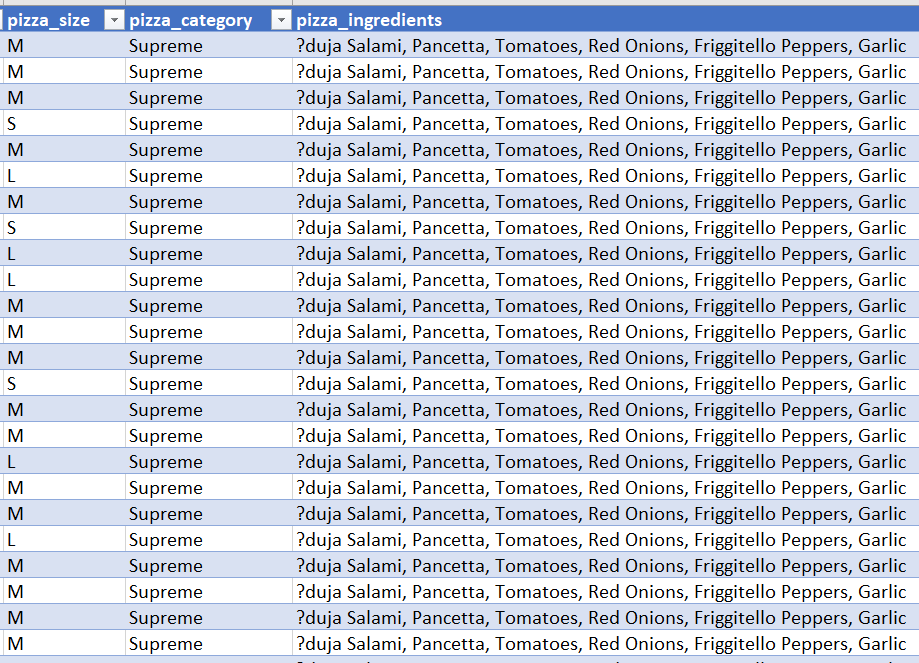
* **Vùng dữ liệu bị trùng lặp:**



**Hình 2.3:** Mô tả lỗi của dữ liệu

Có thể thấy trong bảng order\_id ta thấy rõ ràng rằng số thứ tự của bảng order\_id phải được đánh số theo thứ tự nhưng trong bộ dữ liệu này ta thấy nó bị trùng số 2 trong 5 cột vì vậy ta cần xử lý dữ liệu lại sẽ được thực hiện ở phần sau.

* **Vùng dữ liệu có kí tự đặc biệt hoặc viết sai:**



**Hình 2.4:** Mô tả lỗi của dữ liệu

Ở hình 2.4 như ta thấy tại cột pizza\_ingredients dữ liệu được mô tả trong cột bị thừa kí tự “?duja” lỗi này có thể do lỗi hệ thống xử lý đơn hàng hoặc lỗi phông chữ.

* **Vùng dữ liệu hiện dư thông tin:**



**Hình 2.5:** Mô tả lỗi của dữ liệu

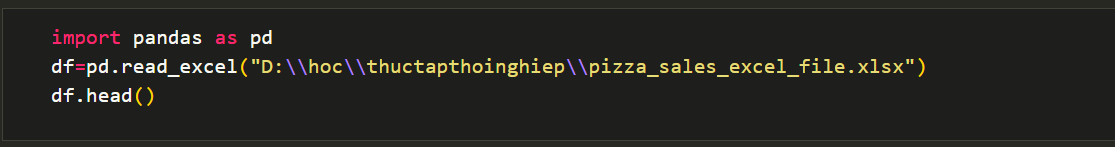
Trong cột pizza\_ingredients ta thấy được rằng nguyên liệu trong cột được mô tả quá nhiều vì thế ta cần tách thành 2 cột để nhìn được trực quan thông tin hơn.

#### **2.2.1.3. Sử dụng Python làm sạch dữ liệu:**

Như đã nói ở phần trên nhóm em sử dụng ngôn ngữ lập trình Python kết hợp với thư viện Pandas và Matplotlib để làm sạch và trực quan hóa dữ liệu sẽ được trình bày qua các bước mô tả như sau:

* Đưa dữ liệu từ Excel vào Python:

Để có thể thao tác với dữ liệu trong Python ta cần đưa dữ liệu từ Excel vào Python bằng cách import thư viện pandas sau đó thực hiện các bước như sau



(3)

(2)

(1)

**Hình 2.6:** Sử dụng thư viện gọi dữ liệu

|  |  |
| --- | --- |
| **STT** | **Mô Tả** |
| (1) | Import thư viện pandas trong Python |
| (2) | Tạo biến gọi DataFrame liên kết đường dẫn tới dữ liệu Excel |
| (3) | Gọi hàm để có thể xuất dữ liệu bảng |

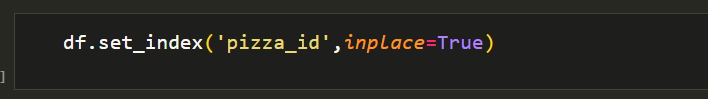
Sau khi thực hiện các bước trên ta sẽ có kết quả như sau:



**Hình 2.7:** Kết quả sau khi đưa dữ liệu vào Python

* Xét giá trị index tìm kiếm dữ liệu bị dư thừa:

Ở bước này như đã nói ở hình 2.4 ta thấy dữ liệu có xuất hiện kí tự lỗi vì vậy sử dụng thư viện Pandas của Python ta có thế xóa toàn bộ dữ liệu dư đó như sau:



**Hình 2.8:** Xét vị trí index cho dữ liệu

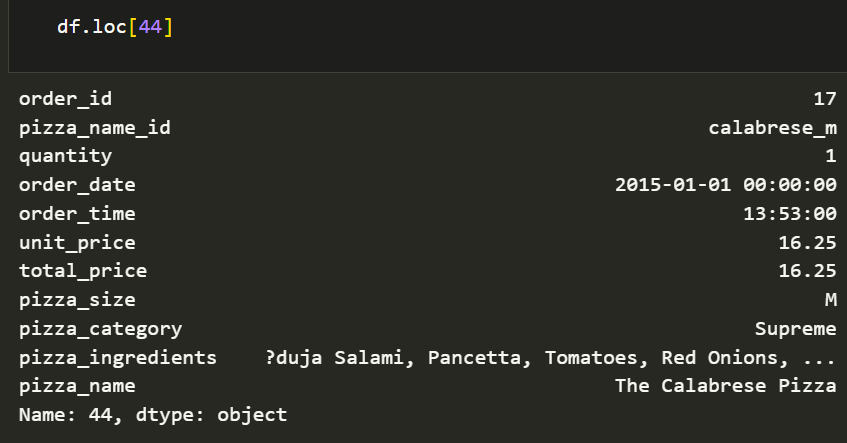
Vì bộ dữ liệu này rất lớn và khó có thể tìm thấy được vì vậy ta cần xét giá trị index cho tập dữ liệu này bằng cách xét cột “pizza\_id” là giá trị index để có thể tìm kiếm dữ liệu chứa kí tự đặc biệt.



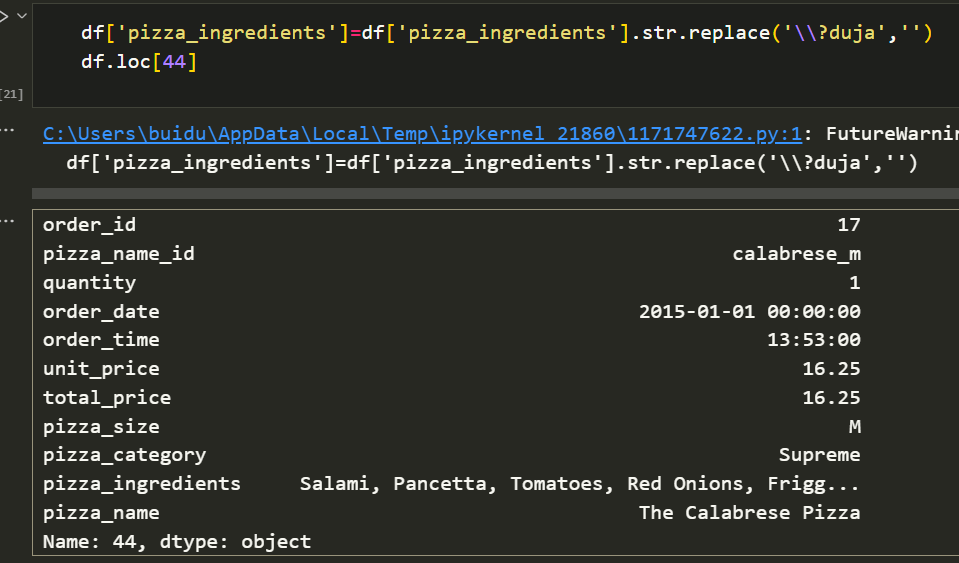
**Hình 2.9:** Kết quả sau khi xét giá trị index

Sau khi xét giá trị index ta có thể thấy kết quả đã có sự thay đổi cột “pizza\_id” đã trở thành cột index vì vậy ta có thể tìm kiếm dữ liệu dễ dàng hơn.

Như vậy hình 2.10 ta thấy được ở cột “pizza\_ingredients” đã xuất hiện dữ liệu có chứa kí tự “?duja” và để loại bỏ kí tự dư đó ta tiếp túc thực hiện gọi hàm trong thư viện Pandas để xử lý dữ liệu thừa đó và xuất ra kết quả như có thể thấy ở hình 2.11.



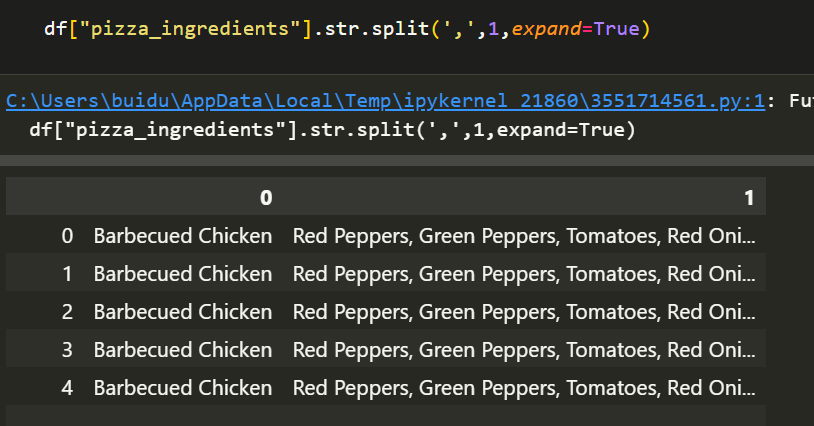
**Hình 2.10:** Dữ liệu dư khi được tìm thấy



**Hình 2.11:** Kết quả sau khi loại bỏ kí tự

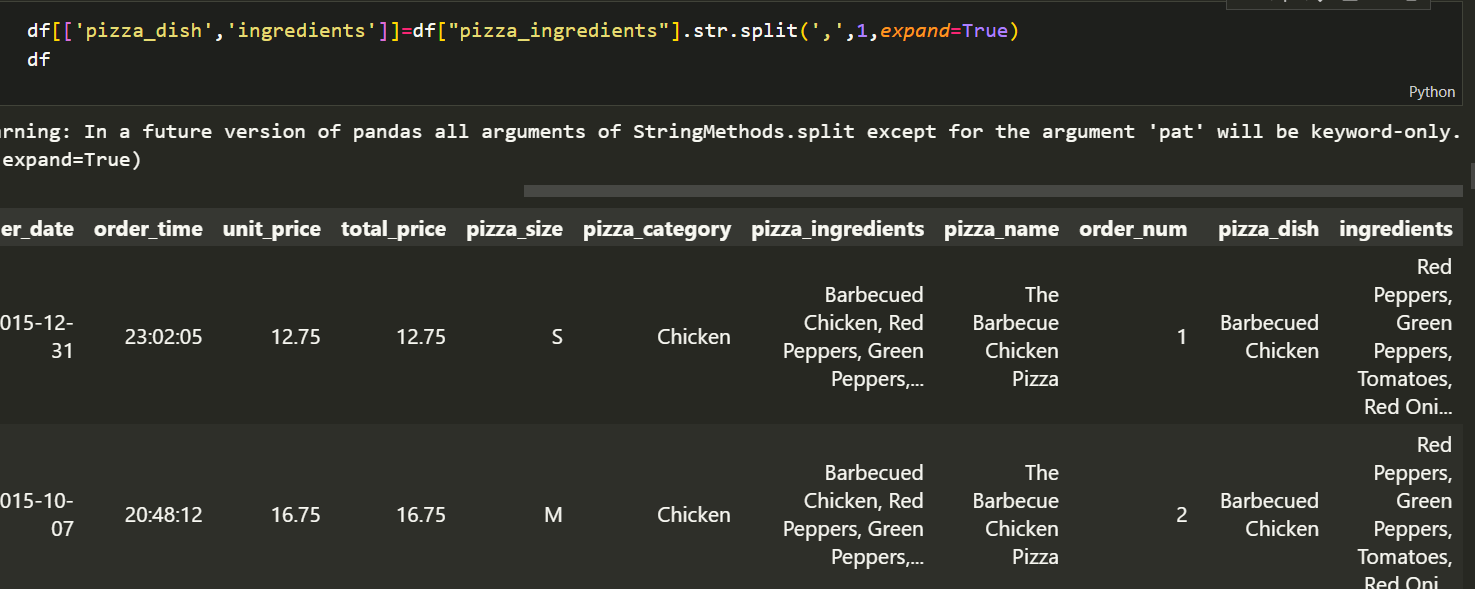
* Tách dữ liệu:

Trong hình 2.2 ở cột “pizza\_ingredients” ta thấy cột có nhiều thông tin dễ gây rối khi nhìn vào vì vậy ta sẽ tách ra làm hai cột giúp ta có thể nhìn dễ dàng hơn.



**Hình 2.12:** Kết quả sau khi tách cột dữ liệu

Sau khi tách dữu liệu ta có thể dễ dàng nhìn hơn nhưng hai cột dữ liệu vẫn chưa được đặt tên nên ta cần đặt tên cho hai cột đó.

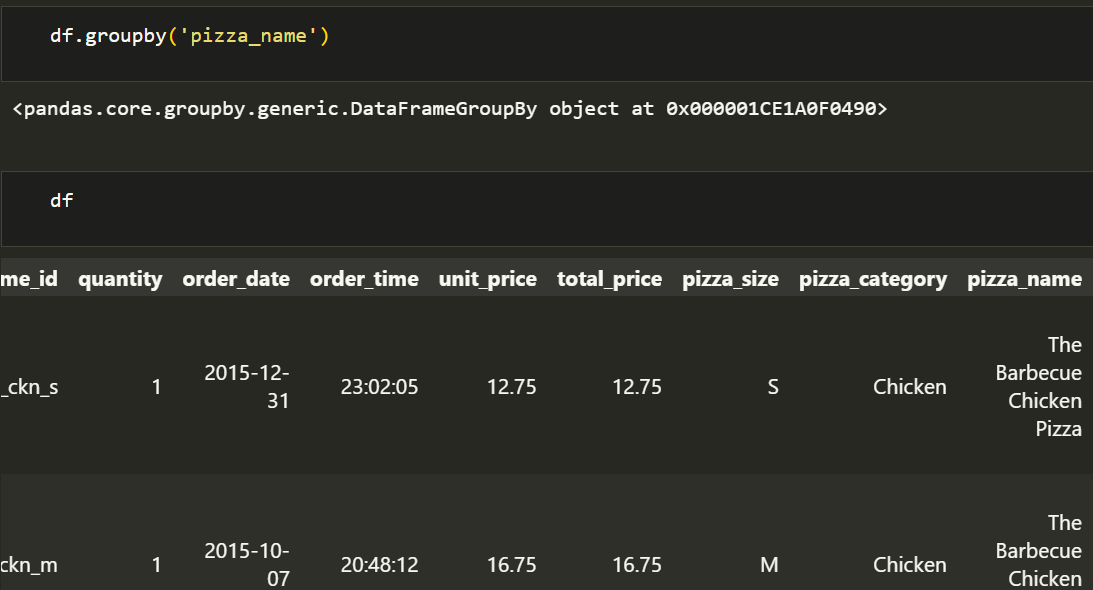


**Hình 2.13:** Kết quả sau khi đặt tên



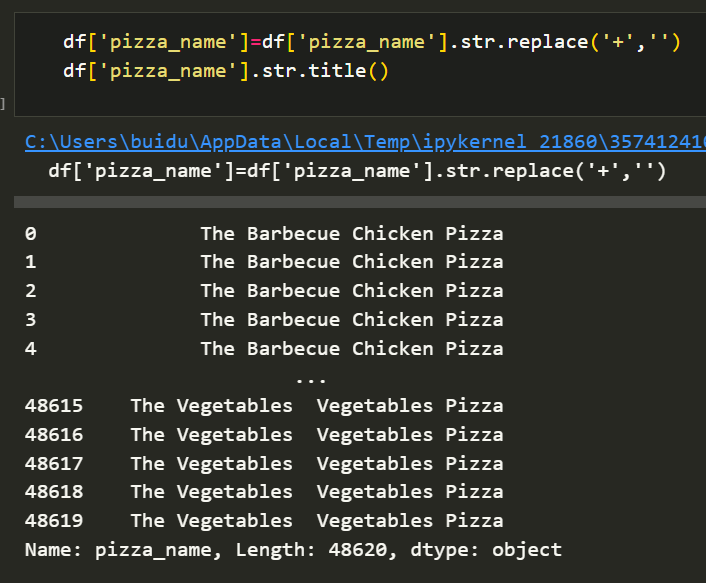
**Hình 2.14:** Xóa cột thừa và xuất kết quả

Để cho dữ liệu được dễ nhìn hơn ta sắp xếp lại dữ liệu bằng cách dùng hàm “groupby” trong pandas.

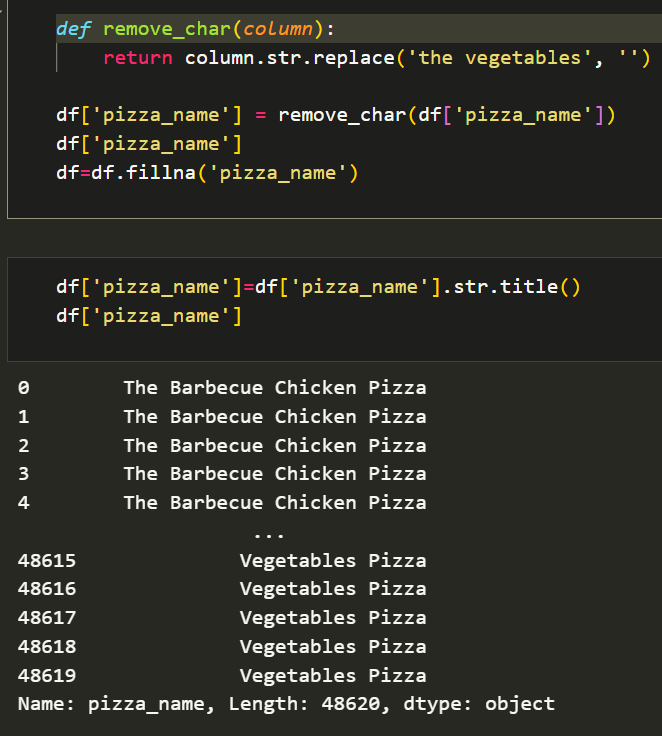


**Hình 2.15:** Dữ liệu sau khi đã sắp xếp

Tương tự như các bước trên ta bỏ đi những kí tự thừa như dấu “+” và xuất kết quả trong tập dữ liệu như hình 2.16, cuối cùng là tạo hàm để loại bỏ dữ liệu “The Vegetables” trong hình 2.16 và xuất kết quả như hình 2.17.



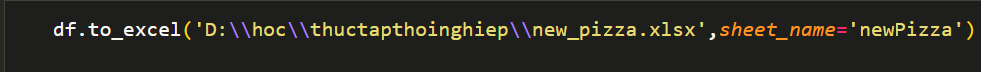
**Hình 2.16:** Kết quá sau khi loại bỏ dữ liệu thừa



**Hình 2.17:** Kết quả sau khi loại bỏ dữ liệu thừa

* Xuất dữ liệu vào file Excel mới:

Sau khi thực hiện xong các bước làm sạch bước cuối cùng ta tạo file Excel mới dùng để chứa dữ liệu đã được xử lý trong Python.



**Hình 2.18:** Xuất kết quả vào file Excel mới

### **2.2.2. Sử dụng SQL thực hiện truy vấn dữ liệu theo yêu cầu:**

Như đề cập ở phần 2.1.2.1 ta tiến hành thực hiện các yêu cầu đề ra để xem xét đánh giá tập dữ liệu chứa thông tin gì quan trọng hay những thông tin gì có giá trị, để từ đó có thể đưa ra nhận định đánh giá khách quan hơn về tập dữ liệu mà nhóm đang sử dụng để phân tích.

|  |  |  |
| --- | --- | --- |
| **Hình ảnh** | **Mô tả** | **Chú thích** |
|  | **Hình 2.19:** Tổng doanh thu của tất cả đơn hàng có trong tập dữ liệu |  |
|  | **Hình 2.20:** Tổng số đơn hàng đã bán được |  |
|  | **Hình 2.21:** Giá trị trung bình cho mỗi đơn hàng |  |
|  | **Hình 2.22:** Trung bình số lượng đơn hàng đã bán qua mỗi đơn hàng |  |
|  | **Hình 2.23:** Tổng số pizza có giá tiền tương ứng | Ở đây ta có thể lựa chọn giá tiền bất kì nếu muốn, ta có thể xem giá tiền của pizza trong tập dữ liệu |
|  | **Hình 2.24:** Tổng số đơn hàng đặt mỗi thứ trong tuần |  |
|  | **Hình 2.25:** Tổng số đơn hàng đặt qua mỗi tháng |  |
|  | **Hình 2.26:** Phần trăm số pizza bán theo loại |  |
|  | **Hình 2.27:** Phần trăm size pizza đã bán trong đơn hàng |  |
|  | **Hình 2.28:** Top 5 pizza có doanh thu thấp nhất |  |
|  | **Hình 2.29:** Top 5 pizza có doanh thu cao nhất |  |
|  | **Hình 2.30:** Top 5 pizza có số lượng đơn bán ra nhiều nhất |  |
|  | **Hình 2.31:** Top 5 pizza có số lượng đơn bán ra ít nhất |  |
|  | **Hình 2.32:** Số pizza bán ra trong một ngày | Trong lệnh truy vấn này đã sử dụng “Store Procedure” để thực hiện truy vấn từng ngày giúp ta biết được cụ thể số bánh đã bán được theo từng ngày |
|  | **Hình 2.33:** Số pizza bán ra được trong một quý | Trong lệnh truy vấn này đã sử dụng “Store Procedure” để thực hiện truy vấn từng ngày giúp ta biết được cụ thể số bánh đã bán được theo từng ngày |
|  | **Hình 2.34:** Số pizza bán ra được trong một tháng |  |
|  | **Hình 2.35:** Số size pizza bán được trong một tháng | Lệnh truy vấn này cho ta biết được size pizza nào bán chạy tùy vào tháng mà ta muốn biết để từ đó đưa ra những chương trình hay ưu đãi nhằm tăng doanh thu. |

### **2.2.4. Mô hình hóa dữ liệu:**

Phần này nhóm sẽ báo cáo một phần riêng do số lượng có hạn của báo cáo.

### **2.2.5. Đưa ra những nhận định, phân tích tổng quát dữ liệu:**

#### **2.2.5.1. Nhận định về tập dữ liệu:**

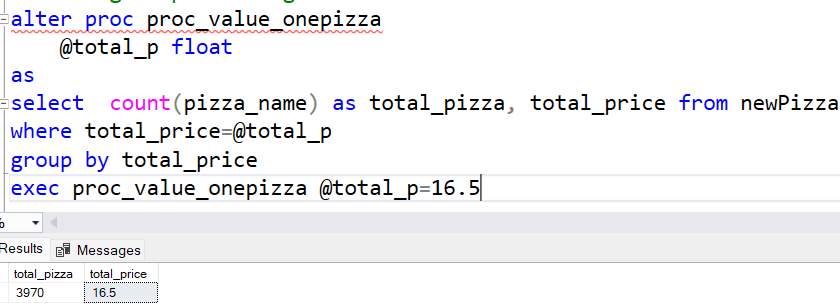
Qua quá trình đưa ra những điều kiện truy vấn, mô hình hóa dữ liệu nhóm đưa ra những dự đoán đánh giá dựa vào những kết quá thu được trong quá trình truy vấn mô hình hóa dữ liệu dựa vào những yếu tố về giá tiền, tên bánh, số lượng, kích thước,… . Dựa vào những yếu tố đó nhóm sẽ đưa ra những nhận định như sau:

* Ở hình 2.21 ta thấy giá trung bình cho mỗi đơn hàng là 16.82 như vậy ta có thể biết được giá trung bình bán ra của một sản phẩm
* Ở hình 2.22 để biết được số lượng đơn hàng đã bán qua mỗi đơn ta lấy tổng số lượng bán ra chia cho số đơn hàng thì ta thấy rằng trung mình mỗi đơn hàng bán ra là 1.02 đơn cho mỗi một giao dịch.
* Để có thể biết được giá trị của từng loại bánh trong tập dữ liệu ta sử dụng “store procedure” hay còn gọi là thủ tục nội tại, nó là tập hợp chứa các dòng lệnh, các biến và các cấu trúc điều khiển trong ngôn ngữ sql, sử dụng quy trình này cho phép đóng gói câu lệnh phực tạp cần gọi hay truy cập thường xuyên, nó giúp cải thiện tốc độ, hiệu suất truy vấn trong quá trình sử dụng[9,tr168]. Vì lệnh gọi giá tiền trong hình 2.23 là khác nhau phụ thuộc vào giá trị mà người dùng muốn biết nên ta có thể nhập bất kì giá tiền nào nếu muốn.
* Hình 2.24 và 2.25 thống kê ra cho thấy số đơn hàng bán ra trong tuần và trong tháng như đã thấy ta có thể biết được thứ mấy hay tháng nào có số lượng đơn đặt hàng nhiều hoặc ít.
* Tương tự như hình 2.26 và 2.27 ta thấy được thống kê tổng giá tiền phần trăm của loại bánh và size bánh bán được.
* Tiếp đến ở hình 2.28 và 2.29 ta thống kê tên bánh pizza có doanh thu cao và thấp nhất dựa vào đó đưa ra những phương án khác nhau nhằm cải thiện hiệu quả kinh doanh.
* Hình 2.30 và 2.31 thấy được thống kê số lượng đơn bán ra nhiều và ít nhất trong tập dữ liệu.
* Hình 2.32 tương tự như hình 2.23 ta cũng sử dụng “store procedure” để truy vấn ngày và tháng cụ thể để có thể xác định được số lượng hàng bán ra trong một ngày.
* Hình 2.33 dựa vào đơn đặt hàng ta có thể xác định được số đơn hàng bán ra trong một quý.

#### **2.2.5.2. Những phương án đưa ra:**

Dựa vào những nhận định đã nói ở trên nhóm sẽ tiến hành đưa ra những phương án, những dự đoán, đưa ra những phương án tối ưu nhất để có thể cải thiện hiệu quả trong quá trình kinh doanh

* Dựa vào giá trị trung bình của mỗi đơn hàng ở hình 2.21 trung bình giá tiền của mỗi đơn hàng bán ra là 16.82 vì vậy khi đưa vào truy vấn lệnh ở hình 2.23 ta thu được kết quả như hình 2.36 điều này chứng tỏ rằng số pizza đã bán nhiều nhất có giá tiền là 16.5 nên phương án đưa ra là cần phát triển hơn loại bánh có giá từ 16.5 trở xuống.



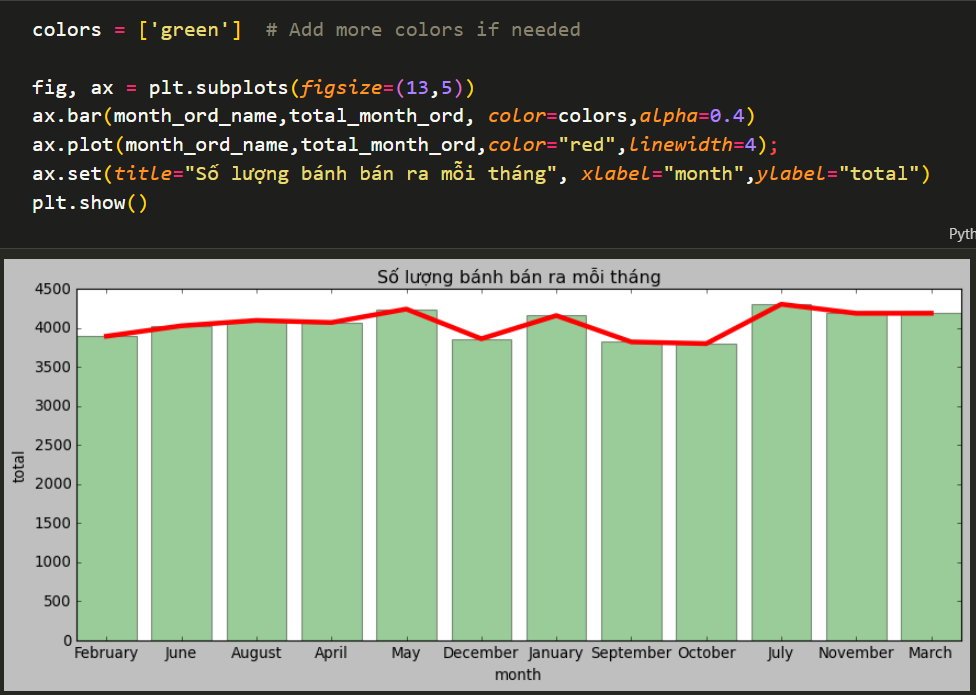
**Hình 2.36:** Kết quả đánh giá

* Như thống kê ở hình 2.24 lượng đơn hàng bán ra trong mỗi thứ cũng có sư chênh lệch nhưng không chênh quá nhiều sử dụng thư viện “Matplotlib” mô hình hóa dữ liệu ta thấy hình 2.37 trong đồ thị ta thấy thứ 6 là ngày bán chạy nhất với số đơn hàng bán ra hơn 8000 đơn trong khi đó ngày chủ nhật là ngày bán được ít đơn hàng nhất, vậy ta cần đưa ra những chương trình khuyến mãi hay giảm giá nhằm kích thích nhu cầu mua hàng của khách hàng.



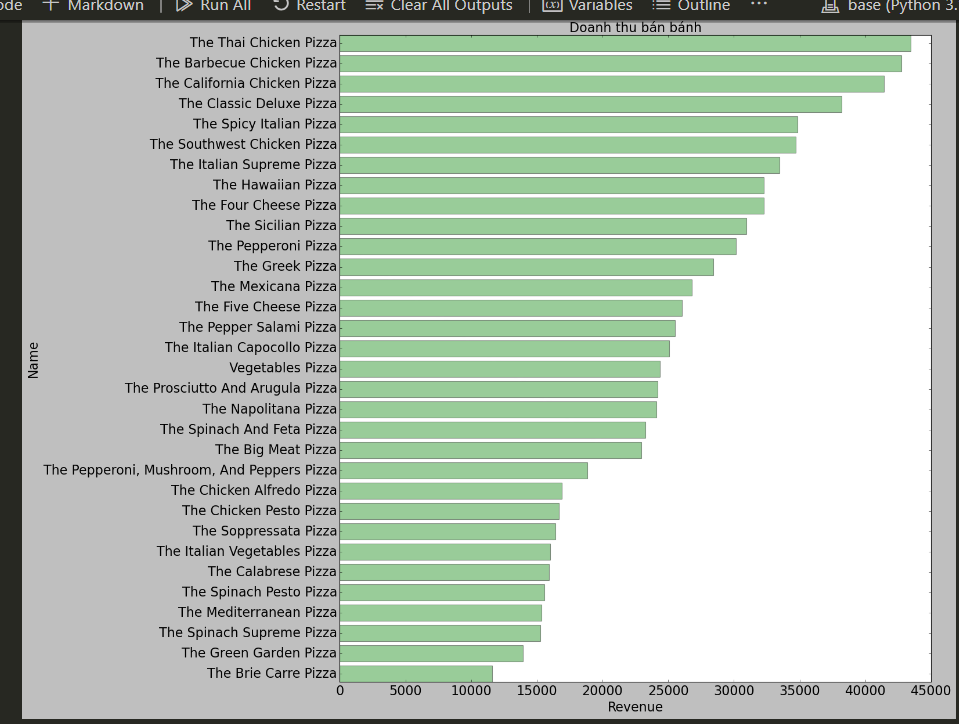
**Hình 2.37:** Đồ thị đánh giá số lượng đơn bán ra

* Hình 2.25 cũng tương tự như vậy với số lượng đơn bán ra mỗi tháng khi mô hình dữ liệu ta thấy được giá bánh bán ra qua mỗi tháng không chênh lệch nhau quá nhiều nên vì thế vẫn ổn định vì vậy ta chưa cần đưa ra phương án để phát triển.



**Hình 2.38:** Mô hình doanh thu qua mỗi tháng

* Hình 2.28 và 2.29 ta thấy tên pizza có doanh thu cao nhất và thấp nhất khi mô hình lên có thể thấy rõ hình 2.39 bánh “The thai chicken pizza” là bánh bán chạy nhất còn bánh “The brie carre pizza” có doanh thu thấp nhất vì vậy phương án cải đưa ra là cần cải thiện bánh có doanh thu thấp hoặc chuẩn bị nguyên liệu ít hơn nhằm hạn chế được sự dư thừa nguyên liệu giảm thiểu tối đa chi phí làm nguyên liệu.



**Hình 2.39:** Mô hình tên bánh theo doanh thu

* Ở hình 2.27 ta thấy doanh thu size pizza trong đơn hàng với size L là có doanh thu cao nhất chứng tỏ rằng size L là size bán chạy nhất chiếm đến 45% trong tập dữ liệu và thấp nhất là size XXL với 0,12% vì thế theo như dự đoán size XXL có giá tiền cao nên ít người có khả năng mua nên ta cần giảm giá hoặc khuyến mãi cho khách hàng thì mới có thể bán nhằm tránh hàng tồn gây lãng phí.

# **III. Kết quả dự án:**

## **3.1. Đánh giá chung:**

Nhìn chung qua quá trình phân tích đưa ra những nhận định về tập dữ liệu trên, nhóm đã đánh giá tổng quát và đưa ra một số nhận định dự đoán về cửa hàng bán bánh pizza. Ta thấy dữ liệu sử dụng để phân tích không có sự chênh lệch về giá khi dựa vào giá trị trung bình ta có thể thấy ngay được giá trung bình mà khách hàng hay mua thường rơi vào khoảng 16 đổ lại.

Về số đơn hàng bán được qua mỗi thứ và mỗi tháng cũng có sự khác biệt, đầu tiên là đơn hàng bán được qua mỗi thứ có thể thấy rằng thứ 6 có đơn hàng cao nhất nhưng khi so với số đơn hàng qua mỗi tháng thì lại không có chệnh lệch nhiều. Cuối cùng là tên bánh pizza bán chạy nhất đã được mô hình hóa ở hình 2.39 theo như đánh giá ta cần tính toán lại tối ưu hơn sản phẩm và các nguyên liệu để có thể cải thiện hơn các sản phẩm mà cửa hàng đang kinh doanh.

Tóm lại những ý tưởng dự đoán đưa ra ở trên ta rút ra một vài điều cần cải thiện như về chất lượng sản phẩm, giảm giá thành, phát triển thêm nhiều chương trình ưu đãi nhằm kích thích khả năng mua hàng của khách hàng từ đó giúp tăng trưởng doanh thu khi kinh doanh.

## **3.2. Những thiếu sót và hạn chế trong quá trình thực hiện đề tài:**

Trong suốt quá trình thực hiện đề tài nhóm rút ra được một số thiếu sót và hạn chế như:

* Dữ liệu sử dụng để phân tích được lấy trên web vẫn còn hạn chế chưa thực tế với dữ liệu ngoài
* Khả năng phân tích còn yếu chưa thực sự tốt, vẫn chưa trực quan và logic.
* Một số nội dung giải thích chưa rõ ràng và mạch lạc vẫn còn thiếu thông tin trong quá trình thực hiện.

## **3.3. Đề xuất phát triển và cải thiện:**

Qua quá trình thực hiện đề tài những vấn đề cần phát triển và cải thiện lại gồm một số vấn đề như sau:

* Về thu thập dữ liệu: Cần sử dụng nguồn dữ liệu có thật hay sát với dữ liệu ngoài đời để có thể phân tích đánh giá trực quan tốt hơn.
* Cần đưa ra những so sánh về dữ liệu so với thực tế, đưa ra nhiều dự đoán sát với nhu cầu thực tế thị trường hơn khi phân tích
* Cải thiện thêm khả năng lập trình, truy vấn và mô hình hóa dữ liệu đưa mô hình dự đoán nhằm dự đoán thị trường chính xác hơn.

Những đề xuất trên là những điều mà nhóm muốn đề cập để có thể cả thiện lại việc phân tích dữ liệu, học tập bổ sung thêm kiến thức để hoàn thiện khả năng nhiều hơn.

# **IV. Kết luận:**

Với mong muốn đánh giá khách quan nhất về tập dữ liệu cửa hàng bán pizza và đưa ra những đánh giá về quá trình kinh doanh thì nhìn chung nhóm đã rút ra được một số thông tin quan trọng và cần thiết, tuy chưa có nhiều kinh nghiệm về mảng phân tích dữ liệu, nhưng nhóm đã cố gắng làm tốt nhất trong khả năng của mình để đưa ra những dự đoán, đánh giá, những thông tin cần thiết nhất với mục đích duy nhất là giúp những nhà điều hành có thể đưa ra những quyết định hay những ý tưởng nhằm nâng cao hiệu quả khi kinh doanh.

# **V. Tài liệu tham khảo:**

1. *Python là gì? - Giải thích về ngôn ngữ Python - AWS (amazon.com)*, <https://aws.amazon.com/vi/what-is/python/>

2. NhiPham, *Xử lý dữ liệu với Pandas trong Python, 26-05-2020*, <https://codelearn.io/sharing/xu-ly-du-lieu-voi-pandas-trong-python>

3. Nguyễn Văn Hoàng, *Giới thiệu về Matplotlib (một thư viện rất hữu ích của Python dùng để vẽ đồ thị), 21/07/2019*, <https://viblo.asia/p/gioi-thieu-ve-matplotlib-mot-thu-vien-rat-huu-ich-cua-python-dung-de-ve-do-thi-yMnKMN6gZ7P>

4. Fabio Nelli. (2015). *Python Data Analytics Data Analysis and Science Using Pandas, matplotlib, and the Python Programming Language*, Springer Science+Business Media, New York, 167-168.

5. *SQL Server là gì? SQL Server giúp bạn làm việc dễ dàng hơn? | TopDev,* <https://topdev.vn/blog/sql-server-la-gi/>

6. *SQL Server Operating System Related Dynamic Management Views (Transact-SQL) - SQL Server | Microsoft Learn,* <https://learn.microsoft.com/en-us/sql/relational-databases/system-dynamic-management-views/sql-server-operating-system-related-dynamic-management-views-transact-sql?view=azuresqldb-current>

7. VinSEP, *Power BI là gì? | Khái niệm và thành phần chính,* <https://vinsep.com/kien-thuc/microsoft-kien-thuc/power-platform/power-bi-la-gi/>

8. *Microsoft Excel – Wikipedia tiếng Việt*, 09/2023, <https://vi.wikipedia.org/wiki/Microsoft_Excel>

9. Stack Overflow contributiors*, Learning SQL*, Creative Commons BY-SA,168.