Shape, rectangle

Description automatically generatedLogo

Description automatically generatedĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

**KHOA HỆ THỐNG THÔNG TIN**



**BÁO CÁO TIỂU LUẬN MÔN HỌC: Kho dữ liệu Olap**

ĐỀ TÀI: Bán hàng trực tuyến AMAZON (2020 – 2021)

Giảng viên hướng dẫn: Nguyễn Thị Kim Phụng

**Sinh viên thực hiện:**

Dương Bảo Tâm - 20521865

Lê Thanh Giang - 20521260

# **LỜI CẢM ƠN**

Nhóm chúng em muốn bày tỏ lòng biết ơn chân thành đến Trường Đại học Công nghệ thông tin - Đại học Quốc gia TP.HCM và khoa Hệ thống thông tin đã cung cấp cho chúng em những kiến thức cơ bản để thực hiện đề tài.

Đặc biệt, chúng em muốn gửi lời cảm ơn sâu sắc tới giảng viên môn Kho dữ liệu, Cô Nguyễn Thị Kim Phụng, vì đã hướng dẫn và đóng góp ý kiến quý báu giúp chúng em hoàn thành tốt báo cáo môn học.

Trong suốt một học kỳ thực hiện đề tài, nhóm chúng em đã áp dụng kiến thức nền tảng và học hỏi kiến thức mới để hoàn thành báo cáo đồ án tốt nhất có thể. Tuy nhiên, chúng em nhận thấy rằng còn tồn tại những thiếu sót, do đó, chúng em rất mong muốn nhận được sự góp ý từ Cô để hoàn thiện kiến thức của mình và chuẩn bị tốt hơn cho các đề tài sắp tới.

Xin chân thành cảm ơn Cô!

**MỤC LỤC**

[**LỜI CẢM ƠN** 2](#_Toc139042945)

[1. GIỚI THIỆU TỔNG QUAN VỀ KHO DỮ LIỆU 4](#_Toc139042946)

[1. Phát biểu về dữ liệu 4](#_Toc139042947)

[1.1 Lý do chọn đề tài: 4](#_Toc139042948)

[1.2 Nguồn kho dữ liệu : 4](#_Toc139042949)

[1.3 Thuộc tính của kho dữ liệu 4](#_Toc139042950)

[2. Xây dựng kho dữ liệu 7](#_Toc139042951)

[2.1 Sơ đồ sao hình minh họa 7](#_Toc139042952)

[2.2 Phân tích dữ liệu từ dữ liệu gốc. 7](#_Toc139042953)

[3. SSIS[1] 11](#_Toc139042954)

[4. SSAS 21](#_Toc139042955)

[5. QUÁ TRÌNH TẠO BÁO CÁO BẰNG CÔNG CỤ POWER BI 32](#_Toc139042956)

[6. Excel Pivot[2] 64](#_Toc139042957)

[7. MDX 83](#_Toc139042958)

[8. DATAMINING[3] 87](#_Toc139042959)

[8.1 Linear Regression 87](#_Toc139042960)

[8.2 Random Forest 102](#_Toc139042961)

[9.REFERENCES 113](#_Toc139042962)

# 

# 1. GIỚI THIỆU TỔNG QUAN VỀ KHO DỮ LIỆU

## Phát biểu về dữ liệu

### Lý do chọn đề tài:

Sau khi tìm hiểu về kho dữ liệu và xem xét nhiều khía cạnh của chủ đề, nhóm đã đưa ra tiêu chí để lựa chọn đề tài cho môn học. Để đáp ứng được yêu cầu của đồ án, đề tài cần phải đáp ứng các tiêu chí sau đây: phải có tính mới lạ, phải có nhiều cột dữ liệu phong phú, cùng với số lượng dòng đủ lớn để thực hiện phân tích. Sau nhiều cuộc họp và tranh luận, nhóm đã chọn đề tài "Bán hàng trực tuyến AMAZON". Đây là lần một đề tài thú vị và đầy tiềm năng, cũng như đáp ứng đủ yêu cầu của nhóm.

### Nguồn kho dữ liệu :

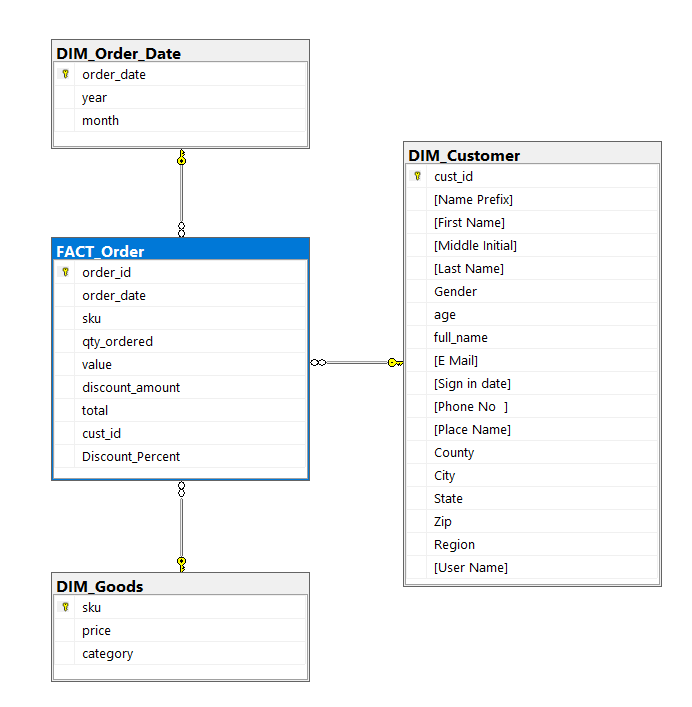
-Link nguồn kho dữ liệu: [Amazon Sales FY2020-21 | Kaggle](https://www.kaggle.com/datasets/earthfromtop/amazon-sales-fy202021)

### Thuộc tính của kho dữ liệu

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Tên thuộc tính** | **Kiểu dữ liệu** | **Mô tả** |
| 1 | order\_id | Int | Mã của đơn hàng |
| 2 | Order\_date | Date | Ngày đặt đơn hàng |
| 3 | Status | Varchar | Trạng thái của đơn hàng |
| 4 | Item\_id | Int | Mã của sản phẩm |
| 5 | Sku | Varchar | Mã SKU của sản phẩm |
| 6 | Qty\_ordered | Int | Số lượng sản phẩm được đặt mua |
| 7 | price | Double | Giá sản phẩm |
| 8 | value | Double | Giá trị tổng của các sản phẩm được đặt mua |
| 9 | Discount\_amount | Double | Số tiền giảm giá được áp dụng cho đơn hàng |
| 10 | total | Double | Tổng số tiền thanh toán cho đơn hàng |
| 11 | category | Varchar | Danh mục sản phẩm |
| 12 | Payment\_method | Varchar | Phương thức thanh toán được sử dụng cho đơn hàng |
| 13 | Bi\_st |  | Giá trị trạng thái của đơn hàng, thể hiện thông tin về tình trạng và giá trị của đơn hàng trong quá trình xử lý và thanh toán |
| 14 | Cust\_id |  | Mã định danh duy nhất cho mỗi khách hàng |
| 15 | year |  | Năm đặt hàng |
| 16 | month |  | Tháng đặt hàng |
| 17 | Name Prefix |  | Tiêu đề hoặc tiền tố tên của khách hàng |
| 18 | First Name |  | Tên của khách hàng |
| 19 | Middle Initial |  | Chữ cái viết tắt của tên đệm của khách hàng |
| 20 | Last Name |  | Họ của khách hàng |
| 21 | Gender |  | Giới tính của khách hàng |
| 22 | Age |  | Tuổi của khách hàng |
| 23 | Full name |  | Họ tên đầy đủ của khách hàng |
| 24 | E Mail |  | Địa chỉ email của khách hàng |
| 25 | Phone No. |  | Số điện thoại của khách hàng |
| 26 | Place Name |  | Tên địa điểm liên quan đến địa chỉ của khách hàng |
| 27 | County |  | Quận liên quan đến chỉ của khách hàng |
| 28 | City |  | Thành phố liên quan đến địa chỉ của khách hàng |
| 29 | State |  | Bang liên quan đến địa chỉ của khách hàng |
| 30 | Zip |  | Mã bưu chính liên quan đến địa chỉ của khách hàng |
| 31 | Region |  | Khu vực liên quan đến địa chỉ của khách hàng |
| 32 | Username |  | Tên đăng nhập của khách hàng |
| 33 | Discount\_Percent |  | Tỷ lệ phần trăm giảm giá được áp dụng cho đơn hàng. |
| 34 | Number ref |  | Số dùng để tham chiếu lên sản phẩm |
| 35 | Sign in date |  | Ngày đăng ký tài khoản của khách hàng |

## Xây dựng kho dữ liệu

### Sơ đồ sao hình minh họa



### Phân tích dữ liệu từ dữ liệu gốc.

-Dữ liệu gốc gồm 36 trường dữ liệu như sau: order\_id,order\_date,status,item\_id,sku,qty\_ordered,price,value,discount\_amount,total,category,payment\_method,bi\_st,cust\_id,year,month,ref\_num,Name Prefix,First Name,Middle Initial,Last Name,Gender,age,full\_name,E Mail,Sign in date,Phone No,PlaceName,County,City,State,Zip,Region,User Name,Discount\_Percent,Paymethod

Sau khi chọn lọc và bỏ đi một số trường dữ liệu không cần thiết thì chúng ta có được bảng DIM và 1 bảng FACT.

**1.Bảng FACT\_Order**

-Mô tả thuộc tính

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Tên thuộc tính** | **Kiểu dữ liệu** | **Mô tả** |
| 1 | order\_id | Int | Mã của đơn hàng |
| 2 | order\_date | Varchar | Ngày đặt đơn hàng |
| 3 | status | double | Mã trạng thái của đơn hàng |
| 4 | sku | Vachar | Mã SKU của sản phẩm |
| 5 | qty\_ordered | Int | Số lượng sản phẩm được đặt mua |
| 6 | value | Double | Giá trị tổng của các sản phẩm được đặt mua |
| 7 | discount\_amount | Double | Số tiền giảm giá được áp dụng cho đơn hàng |
| 8 | bi\_st\_id | Int | Mã giá trị trạng thái của đơn hàng, thể hiện thông tin về tình trạng và giá trị của đơn hàng trong quá trình xử lý và thanh toán |
| 9 | cust\_id | Int | Mã khách hàng |
| 10 | discount\_percent | Double | Tổng số tiền thanh toán cho đơn hàng |
| 11 | total | Double | Tỷ lệ phần trăm giảm giá được áp dụng cho đơn hàng. |

**2.Bảng DIM\_Order\_Date**

-Mô tả thuộc tính

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Tên thuộc tính** | **Kiểu dữ liệu** | **Mô tả** |
| 1 | order\_date | date | Ngày đặt đơn hàng |
| 2 | month | date | Năm đặt đơn hàng |
| 3 | year | date | Tháng đặt đơn hàng |

**3.Bảng DIM\_Goods**

-Mô tả thuộc tính

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Tên thuộc tính** | **Kiểu dữ liệu** | **Mô tả** |
| 1 | sku | Int | Mã SKU của sản phẩm |
| 2 | category\_id | Varchar | Mã danh mục sản phẩm |
| 3 | price | double | Đơn giá của sản phẩm |

**4.Bảng DIM\_Customer**

-Mô tả thuộc tính

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Tên thuộc tính** | **Kiểu dữ liệu** | **Mô tả** |
| 1 | cust\_id | Int | Mã định danh duy nhất cho mỗi khách hàng |
| 2 | full\_name | Varchar | Họ tên đầy đủ của khách hàng |
| 3 | gender | Varchar | Giới tính của khách hàng  -F: là nữ  -M: là nam |
| 4 | age | Int | Tuổi của khách hàng |
| 5 | e\_mail | Varchar | Địa chỉ email của khách hàng |
| 6 | Name Prefix | Varchar | Tiêu đề hoặc tiền tố tên của khách hàng |
| 7 | phone\_no | Varchar | Số điện thoại của khách hàng |
| 8 | zip | Int | Mã bưu chính liên quan đến địa chỉ của khách hàng |
| 9 | city | Varchar | Thành phố liên quan đến địa chỉ của khách hàng |
| 10 | county | Varchar | Quận liên quan đến chỉ của khách hàng |
| 11 | region | Varchar | Khu vực liên quan đến địa chỉ của khách hàng |
| 12 | place\_name | Varchar | Tên địa điểm liên quan đến địa chỉ của khách hàng |
| 13 | first\_name | Varchar | Tên của khách hàng |
| 14 | Last Name | Varchar | Họ của khách hàng |
| 15 | Middle Initial | Varchar | Chữ cái viết tắt của tên đệm của khách hàng |
| 16 | State | Varchar | Bang liên quan đến địa chỉ của khách hàng |
| 17 | Username | Varchar | Tên đăng nhập của khách hàng |
| 18 | Sign in date | Varchar | Ngày đăng ký tài khoản của khách hàng |

# SSIS[1]

Bước 1 : Tạo Data flow để dẫn data từ dataset gốc

A screenshot of a computer

Description automatically generated

Bước 2 : Chọn file dataset gốc từ excel

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Bước 3: Kéo thả các toolbox từ bảng vào để chuẩn bị cho các bước chuyển đổi data từ excel sang sql server

A screenshot of a computer program

Description automatically generated with medium confidence

Bước 4: Nối các liên kết

A screenshot of a computer

Description automatically generated

Bước 5 : Chọn khóa chính ở sort để chuẩn bị tạo bảng

A screenshot of a computer

Description automatically generated

Bước 6 : Tạo một database mới

A screenshot of a computer

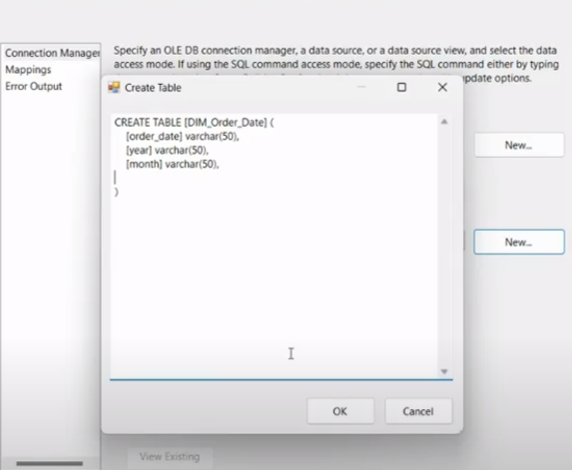
Description automatically generated with medium confidence

Bước 7: Kết nối với server của máy và chọn DB đã tạo sẵn ở tool box OLE DB Destination

A screenshot of a computer

Description automatically generated

Bước 8: Sau đó tạo bảng từ dữ liệu gốc được đổ vào từ file excel bằng cách chọn new table



Sau khi hoàn thành các bước trên ta sẽ có được hình như sau:

A picture containing text, screenshot, rectangle, font

Description automatically generated

Bước 9: Tiếp tục ở tag control flow chọn toolbox Execute SQL task, sau đó nối qua data flow task như hình dưới

A screenshot of a computer

Description automatically generated with low confidence

Bước 10: Nhấp chuột phải để edit SQL Task , chọn phần connection là server của máy và phần sql statement là các câu lệnh xóa đi các bảng ở trong data flow

A screenshot of a computer

Description automatically generated

Bước 11: Chạy chương trình, nếu như không có lỗi chương trình sẽ có dạng như hình bên dưới

A picture containing text, screenshot, line, font

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

Bước 12 : Kiểm thử data ở sql server

A screenshot of a computer

Description automatically generated

# SSAS

Bước 1: Chọn new data source ở phần data source

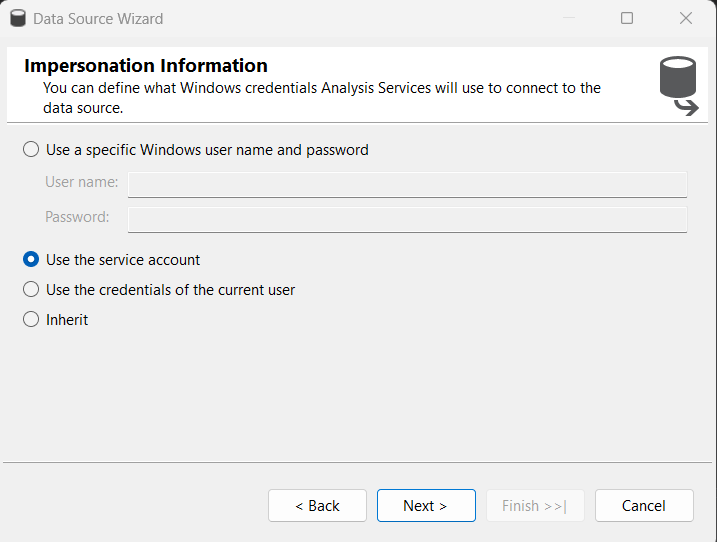
A screenshot of a computer

Description automatically generated

Bước 2 : Nhập vào server của máy và điền vào các thông tin cần thiết, sau đó chọn database



Bước 3 : Sau khi ấn next chọn phần use the service account

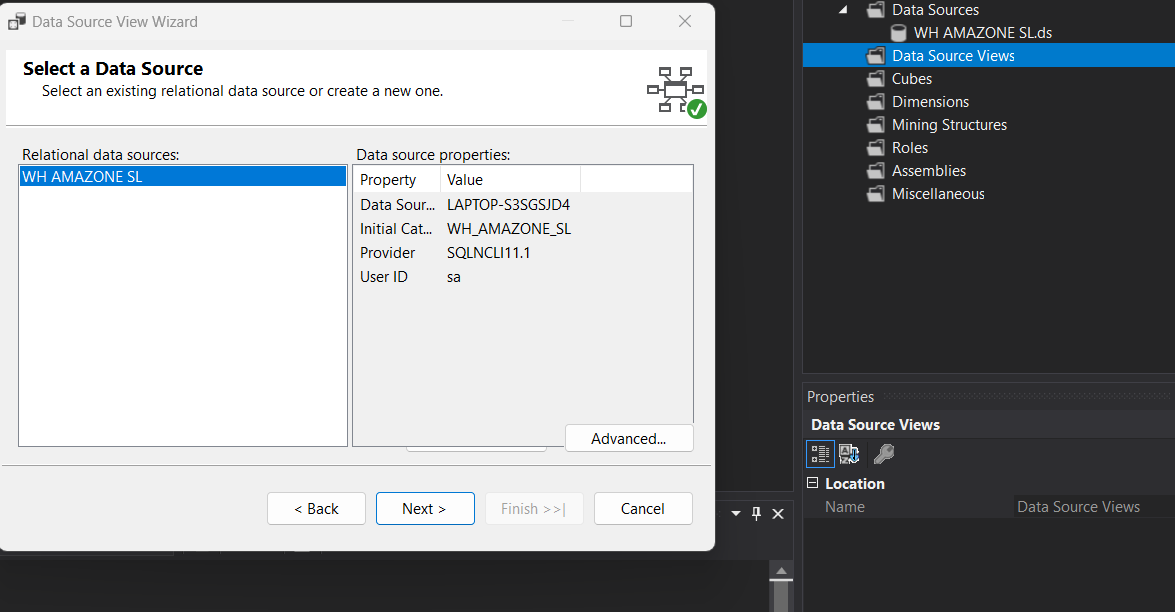


Sau đó ấn finish

A screenshot of a computer

Description automatically generated with medium confidence

Bước 4 : Tạo new data source view ở phần data source view



Nhấn next và chọn đẩy bảng FACT sang phần included object bằng nút > , sau đó ấn add related table , nếu như các table đã có liên kết với nhau thì sẽ tử động chuyển sang bên phần included object

A screenshot of a computer

Description automatically generated

Sau đó ấn finish để hoàn thành quá trình

A screenshot of a computer

Description automatically generated with medium confidence

Bước 5: Tạo cubes mới ở phần cubes

A screenshot of a computer

Description automatically generated

Sau đó ấn next , chọn bảng Fact ở phần measure group table

A screenshot of a computer

Description automatically generated with low confidence

A screenshot of a computer

Description automatically generated with low confidence

Sau đó ấn next và chọn các bảng còn lại ở phần DIM

A screenshot of a computer

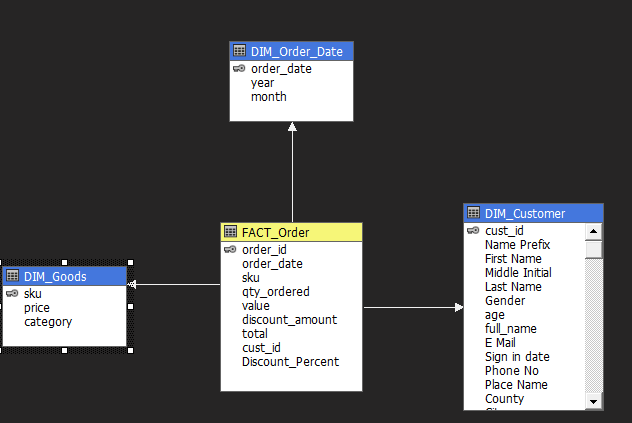
Description automatically generated with medium confidence

Sau đó là ấn Finished

A screenshot of a computer

Description automatically generated with medium confidence

Khi finish sẽ tạo ra được hình như bên dưới



# 5. QUÁ TRÌNH TẠO BÁO CÁO BẰNG CÔNG CỤ POWER BI

5.1 Nhập dữ liệu nguồn (Data Source)

Bước 1: Mở Power BI Desktop, tại mục Home > Get Data > Analysis Service để import dữ liệu

A screenshot of a computer

Description automatically generated

Sau đó chọn mục excel workbook và import dữ liệu từ file excel vào

5.2 Các bước thực hiện

5.2.1 Đếm số lượng mã đơn hàng (order\_id) ở thành phố BrownsTown (City) theo từng năm

Bước 1 :Tạo header

**Đếm số lượng mã đơn hàng (order\_id) ở thành phố BrownsTown (City) theo từng năm**

Bước 2 :Tại cửa sổ visualization chọn Cluster Column Chart

A screenshot of a computer

Description automatically generated

Bước 3: Tại cửa sổ Fields > Chọn các thuộc tính cần truy vấn (Count of order\_id, city,year)

A screenshot of a computer

Description automatically generated with medium confidence

Bước 4: Sau khi đã kéo thả các thuộc tính cần thiết vào biểu đồ thì biểu đồ xuất ra kết quả:

A picture containing text, screenshot, diagram, software

Description automatically generated

Hình 5.1

Biểu đồ cột liên cụm thể hiện số mã đơn hàng ở thành phố Brownstown qua từng năm

A picture containing text, screenshot, font, line

Description automatically generated

Hình 5.2

Ma trận thể hiện số mã đơn hàng ở thành phố Brownstown qua từng năm

5.2.2 Cho biết khu vực (Region) có số đơn hàng nhiều nhất và ít nhất từ tháng 10 2020 đến tháng 9 2021

Bước 1 :Tạo header

**Cho biết khu vực (Region) có số đơn hàng nhiều nhất và ít nhất từ tháng 10 2020 đến tháng 9 2021**

Bước 2 :Tại cửa sổ visualization chọn Stack Column Chart

A screenshot of a computer

Description automatically generated

Bước 3: Tại cửa sổ Fields > Chọn các thuộc tính cần truy vấn (Count of order\_id,Region)

A screenshot of a computer

Description automatically generated with medium confidence

Bước 4: Sau khi đã kéo thả các thuộc tính cần thiết vào biểu đồ thì biểu đồ xuất ra kết quả:

A screenshot of a graph

Description automatically generated with low confidence

Hình 5.3

Biểu đồ cột chồng thể hiện số mã đơn hàng ở 2 khu vực nhiều nhất và ít nhất

A picture containing text, font, screenshot, line

Description automatically generated

Hình 5.4

Ma trận thể hiện số mã đơn ở 2 khu vực nhiều nhất và ít nhất

5.2.3 Cho biết 5 thành phố có nhiều đơn hàng nhất trong cả năm 2021

Bước 1 :Tạo header

**Cho biết 5 thành phố có nhiều đơn hàng nhất trong cả 2020 và 2021**

Bước 2 :Tại cửa sổ visualization chọn Clustered bar chart và Filled map

A screenshot of a computer

Description automatically generated A screenshot of a computer

Description automatically generated

Bước 3: Tại cửa sổ Fields > Chọn các thuộc tính cần truy vấn (Count of order\_id, city,)

A screenshot of a computer

Description automatically generated with medium confidence A screenshot of a phone

Description automatically generated with medium confidence

Bước 4: Tại cửa sổ Filters chọn City sau đó click nhấn mũi tên, sau đó chọn filter type Top N > sau đó chọn Top để lọc ra các giá trị lớn nhất và điền vào số 5 để lọc >kéo thả phần Order\_Id từ bảng Data>chọn Count(Distinct)>sau đó nhấn apply filter

A screenshot of a computer

Description automatically generated with medium confidence

Bước 5: Sau khi đã kéo thả các thuộc tính cần thiết vào biểu đồ thì biểu đồ xuất ra kết quả:

A close-up of a map

Description automatically generated with medium confidence

Hình 5.5

Biểu đồ cột liên cụm thể hiện số mã đơn hàng ở các thành phố và vị trí của các thành phố năm 2021

A picture containing text, screenshot, plot, line

Description automatically generated

Hình 5.6

Biểu đồ cột liên cụm thể hiện số mã đơn hàng ở các thành phố trong năm 2021

A screenshot of a computer

Description automatically generated with low confidence

Hình 5.7

Ma trận thể hiện số mã đơn ở các thành phố

A map of united states

Description automatically generated with medium confidence

Hình 5.8

Vị trí của các thành phố

A screenshot of a computer

Description automatically generated with low confidence

Hình 5.9

Ma trận thể hiện số mã đơn ở các thành phố

5.2.4 Hãy cho biết 3 thành phố có số lượng đơn hàng thấp nhất và tổng chi phí đơn hàng của thành phố trong năm 2020

Bước 1 :Tạo header

**Hãy cho biết 3 thành phố có số lượng đơn hàng thấp nhất và tổng chi phí đơn hàng của thành phố trong năm 2020**

Bước 2 :Tại cửa sổ visualization chọn Pie Chart

A screenshot of a computer

Description automatically generated

Bước 3: Tại cửa sổ Fields > Chọn các thuộc tính cần truy vấn (Sum of value,Sum of qty ordered, city,)

A screenshot of a computer

Description automatically generated with medium confidence A screenshot of a computer

Description automatically generated with medium confidence

Bước 4: Tại cửa sổ Filters chọn City sau đó click nhấn mũi tên, sau đó chọn filter type Top N > sau đó chọn Bottom để lọc ra các giá trị lớn nhất và điền vào số 3 để lọc > kéo thả phần qty\_Order từ bảng Data>sau đó nhấn apply filter >kéo thả thêm một phần City từ bảng data > chọn filter type advanced filtering>chọn start with ở phần show items>gõ vào chữ W (để lấy các giá trị thấp nhất sau cùng theo bảng chữ cái)

A screenshot of a computer

Description automatically generated with medium confidence

Bước 5: Sau khi đã kéo thả các thuộc tính cần thiết vào biểu đồ thì biểu đồ xuất ra kết quả:

A screenshot of a graph

Description automatically generated with low confidence

Hình 5.10

Biểu đồ tròn thể hiện số tổng số đơn hàng và tổng giá tiền đơn hàng của 3 thành phố thấp nhất trong năm 2020

A picture containing screenshot, diagram, text, colorfulness

Description automatically generated

Hình 5.10

Biểu đồ tròn thể hiện số tổng số đơn hàng của 3 thành phố thấp nhất trong năm 2020

A screenshot of a computer

Description automatically generated with low confidence

Hình 5.11

Ma trận thể hiện số tổng số đơn hàng và tổng giá tiền đơn hàng của 3 thành phố thấp nhất trong năm 2020

A picture containing screenshot, text, diagram, font

Description automatically generated

Hình 5.12

Biểu đồ tròn thể hiện số tổng giá tiền đơn hàng của 3 thành phố thấp nhất trong năm 2020

A screenshot of a computer

Description automatically generated with low confidence

Hình 5.13

Ma trận thể hiện số tổng số đơn hàng và tổng giá tiền đơn hàng của 3 thành phố thấp nhất trong năm 2020

5.2.5 Với từng loại sản phẩm hãy liệt kê doanh thu của bang RI (State) trong năm 2020

Bước 1 :Tạo header

**Với từng loại sản phẩm hãy liệt kê doanh thu của bang RI (State) trong năm 2020**

Bước 2 :Tại cửa sổ visualization chọn Pie Chart

A screenshot of a computer

Description automatically generated

Bước 3: Tại cửa sổ Fields > Chọn các thuộc tính cần truy vấn (Sumofvalue,State,Category,Order\_date,Month)

A screenshot of a computer

Description automatically generated with medium confidence

Bước 4: Sau khi đã kéo thả các thuộc tính cần thiết vào biểu đồ thì biểu đồ xuất ra kết quả:

A screenshot of a computer

Description automatically generated with low confidence

Hình 5.14

Biểu đồ tròn thể hiện số tổng doanh thu của bang Ri trong năm 2020

A picture containing text, screenshot, font, number

Description automatically generated

Hình 5.15

Ma trận thể hiện số tổng số doanh thu của bang Ri trong năm 2020

5.2.6 Chọn ra 10 đơn hàng có giá trị lớn nhất tại hạt (County) Accomack theo từng tháng trong năm 2020

Bước 1 :Tạo header

**Chọn ra 10 đơn hàng có giá trị lớn nhất tại hạt (County) Accomack theo từng tháng trong năm 2020**

Bước 2 :Tại cửa sổ visualization chọn Cluster Column Chart

A screenshot of a computer

Description automatically generated

Bước 3: Tại cửa sổ Fields > Chọn các thuộc tính cần truy vấn (Sumofvalue,County,Year,Month)

A screenshot of a computer

Description automatically generated with medium confidence

Bước 4: Sau khi đã kéo thả các thuộc tính cần thiết vào biểu đồ thì biểu đồ xuất ra kết quả:

A screenshot of a graph

Description automatically generated

Hình 5.16

Biểu đồ cột thể hiện số tổng doanh thu của các tháng trong hạt Accomack trong năm 2020

A picture containing text, font, screenshot, number

Description automatically generated

Hình 5.17

Ma trận thể hiện số tổng số doanh thu của các tháng trong hạt Accomack trong năm 2020

5.2.7 Liệt kê 5 gmail có số đơn hàng nhiều nhất trong tháng 10 năm 2020.

Bước 1 :Tạo header

**Liệt kê 5 gmail có số đơn hàng nhiều nhất trong tháng 10 năm 2020.**

Bước 2 :Tại cửa sổ visualization chọn Cluster Column Chart

A screenshot of a computer

Description automatically generated

Bước 3: Tại cửa sổ Fields > Chọn các thuộc tính cần truy vấn (Count of order\_id,Order\_Date,Year,Month,email)

A screenshot of a computer

Description automatically generated with medium confidence

Bước 4: Tại cửa sổ Filters chọn E Mail sau đó click nhấn mũi tên, sau đó chọn filter type Top N > sau đó chọn Top để lọc ra các giá trị lớn nhất và điền vào số 5 để lọc > kéo thả phần Order\_Id từ bảng Data>chọn Count(Distinct )>sau đó nhấn apply filter

A screenshot of a computer

Description automatically generated with medium confidence

Bước 5: Sau khi đã kéo thả các thuộc tính cần thiết vào biểu đồ thì biểu đồ xuất ra kết quả:

A screenshot of a computer

Description automatically generated with medium confidence

Hình 5.18

Biểu đồ cột thể hiện 5 email có số đơn hàng nhiều nhất trong tháng 10 năm 2020

A screenshot of a computer

Description automatically generated with medium confidence

Hình 5.19

Ma trận thể hiện5 email có số đơn hàng nhiều nhất trong tháng 10 năm 2020

5.2.8 Hãy cho biết 4 bang có tổng số đơn hàng thấp nhất trong năm 2020

Bước 1 :Tạo header

**Hãy cho biết 4 bang có tổng số đơn hàng thấp nhất trong năm 2020**

Bước 2 :Tại cửa sổ visualization chọn Stack Column Chart và Filled map

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

Bước 3: Tại cửa sổ Fields > Chọn các thuộc tính cần truy vấn (Count of order\_id,Order\_Date,Year,State)

A screenshot of a computer

Description automatically generated with medium confidence A screenshot of a computer

Description automatically generated with medium confidence

Bước 4: Tại cửa sổ Filters chọn State sau đó click nhấn mũi tên, sau đó chọn filter type Top N > sau đó chọn Bottom để lọc ra các giá trị lớn nhất và điền vào số 3 để lọc > kéo thả phần Order\_id từ bảng Data>sau đó nhấn apply filter >kéo thả thêm một phần State từ bảng data > chọn filter type advanced filtering>chọn does not start with ở phần show items>gõ vào chữ A (để tên của các bang lấy các giá sau cùng theo bảng chữ cái)> tiếp tục chọn does not contain ở dưới > gõ vào chữ J (để tên của các bang lấy các giá sau cùng theo bảng chữ cái)

A screenshot of a phone

Description automatically generated with medium confidence

Bước 5: Sau khi đã kéo thả các thuộc tính cần thiết vào biểu đồ thì biểu đồ xuất ra kết quả:

A map of the united states

Description automatically generated with medium confidence

Hình 5.20

Vị trí của các bang

A screenshot of a computer

Description automatically generated with low confidence

Hình 5.21

Ma trận thể hiện số mã đơn ở các bang

A picture containing screenshot, text, diagram, plot

Description automatically generated

Hình 5.22

Biểu đồ cột chồng thể hiện số mã đơn hàng ở 4 bang

A screenshot of a computer

Description automatically generated with low confidence

Hình 5.23

Ma trận thể hiện số mã đơn ở các bang

A screenshot of a graph

Description automatically generated with low confidence

Hình 5.23

Biểu đồ cột liên thể hiện số mã đơn hàng ở các thành phố và vị trí của các thành phố trong năm 2020

5.2.9 Hãy cho biết tổng chi phí của các đơn hàng ở thành phố Alto theo từng tháng, từng năm

Bước 1: Tạo Header

**Hãy cho biết tổng chi phí của các đơn hàng ở thành phố Alto theo từng tháng, từng năm**

Bước 2 :Tại cửa sổ visualization chọn Cluster Column Chart

A screenshot of a computer

Description automatically generated

Bước 3: Tại cửa sổ Fields > Chọn các thuộc tính cần truy vấn (Sum of total, city,year,month)

A screenshot of a computer

Description automatically generated with low confidence

Bước 4: Sau khi đã kéo thả các thuộc tính cần thiết vào biểu đồ thì biểu đồ xuất ra kết quả:

A screenshot of a graph

Description automatically generated

Hình 5.24

Biểu đồ cột chồng thể hiện tổng doanh thu trong tháng 10 năm 2020 của thành phố Alto

A screenshot of a computer

Description automatically generated with low confidence

Hình 5.25

Ma trận thể hiện tổng doanh thu trong tháng 10 năm 2020 của thành phố Alto

5.2.10 Hiển thị số lượng đặt hàng của độ tuổi 70 trong mỗi tháng của năm 2020

Bước 1 :Tạo header

**Hiển thị số lượng đặt hàng của độ tuổi 70 trong mỗi tháng của năm 2020**

Bước 2 :Tại cửa sổ visualization chọn Cluster Column Chart

A screenshot of a computer

Description automatically generated

Bước 3: Tại cửa sổ Fields > Chọn các thuộc tính cần truy vấn (Sum of total, city,year,month)

A screenshot of a computer

Description automatically generated with medium confidence

Bước 4: Sau khi đã kéo thả các thuộc tính cần thiết vào biểu đồ thì biểu đồ xuất ra kết quả:

A screenshot of a computer

Description automatically generated with medium confidence

Hình 5.26

Biểu đồ cột chồng thể hiện tổng số đơn hàng mà độ tuổi 70 đã mua trong năm 2020

A screenshot of a computer

Description automatically generated with low confidence

Hình 5.27

Ma trận thể hiện tổng số đơn hàng mà độ tuổi 70 đã mua trong năm 2020

# Excel Pivot[2]

Bước 1 : chọn một ô bất kì trong file excel chứa dữ liệu

A screenshot of a computer

Description automatically generated

Bước 2 :Nhấn vào mục Insert chọn Table > Ấn ok

A screenshot of a computer

Description automatically generated with medium confidence

Sẽ được hình như bên dưới

A screenshot of a computer

Description automatically generated

Bước 3: Tiếp tục vào mục insert chọn Pivot Table > chọn New Worksheet >chọn add this data to the data model>nhấn ok

A screenshot of a computer

Description automatically generated

Màn hình sẽ xuất hiện hình như sau

A picture containing screenshot, text, plot

Description automatically generated

Bước 4: Kéo thả cho câu truy vấn

Câu 1: Đếm số lượng mã đơn hàng (order\_id) ở thành phố BrownsTown (City) theo từng năm

Kéo thả các mục cần thiết từ bảng Fields( City,Order\_date(year),order\_id)

Để xuất hiện mục đếm lọc trùng thì ta click chuột vào order\_id ở values, chọn Value Field settings> chọn distinct count

A screenshot of a computer

Description automatically generated with medium confidence

Sẽ được hình như sau

A screenshot of a computer

Description automatically generated

Chọn Brownstown ở phần filter city



Kết quả cuối cùng sẽ được như hình bên dưới

A screenshot of a graph

Description automatically generated with medium confidence

Câu 2: Cho biết khu vực (Region) có số đơn hàng nhiều nhất và ít nhất từ tháng 10-2020 đến tháng 9-2021

Bước 1:

Kéo thả các mục cần thiết từ bảng Fields( City,Order\_date(year), Order\_date(month)order\_id)

Để xuất hiện mục đếm lọc trùng thì ta click chuột vào order\_id ở values, chọn Value Field settings> chọn distinct count

A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a computer program

Description automatically generated with medium confidence

Sau khi chọn sẽ ra được kết quả như sau:

A screenshot of a computer

Description automatically generated with medium confidence

Vì yêu cầu là lấy 2 Region cao nhất nên ở phần filter bỏ đi 2 mục là South và West

Đáp án cuối cùng

A screenshot of a computer

Description automatically generated with low confidence

Câu 3: Cho biết 5 thành phố có nhiều đơn hàng nhất trong năm 2021

Bước 1:

Kéo thả các mục cần thiết từ bảng Fields( City,Order\_date(year), order\_id)

Để xuất hiện mục đếm lọc trùng thì ta click chuột vào order\_id ở values, chọn Value Field settings> chọn distinct count

A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated

Ở phàn year fliter chọn 2021



Bước 2:

Nhấn vào nút autosort



Chọn Value filters > chọn mục top10 và chỉnh sửa theo hình bên dưới

A screenshot of a computer

Description automatically generated with medium confidence

Kết quả cuối cùng sẽ được như hình bên dưới

A screenshot of a graph

Description automatically generated with medium confidence

Câu 4: Hãy cho biết 3 thành phố có số lượng đơn hàng thấp nhất và tổng chi phí đơn hàng của thành phố trong năm 2020

Bước 1:

Kéo thả các mục cần thiết từ bảng Fields( City,Order\_date(year), qty\_ordered,value)

A screenshot of a computer

Description automatically generated

Ở phàn year fliter chọn 2020



Bước 2 :

Chọn phần auto sort



Chọn Value filters > chọn mục top10 và chỉnh sửa theo hình bên dưới

A screenshot of a computer

Description automatically generated with medium confidence

Kết quả cuối cùng sẽ được như hình sau:

A screenshot of a spreadsheet

Description automatically generated with medium confidence

Câu 5: Với từng loại sản phẩm hãy liệt kê doanh thu của bang RI (State) trong năm 2020

Bước 1:

Kéo thả các mục cần thiết từ bảng Fields( Category,State, value)

A screenshot of a computer

Description automatically generated

Sau đó ở phần state filter chọn Ri



Kết quả cuối cùng sẽ được như hình bên dưới

A screenshot of a graph

Description automatically generated with medium confidence

Câu 6: Chọn ra 10 đơn hàng có giá trị lớn nhất tại hạt (County) Accomack theo từng tháng trong năm 2020

Bước 1:

Kéo thả các mục cần thiết từ bảng Fields( order\_date(year),order\_date(month),County, value)

A screenshot of a computer

Description automatically generated

Sau đó ở phần County filter chọn Accomack



Kết quả cuối cùng sẽ giống như hình sau

A screenshot of a graph

Description automatically generated with medium confidence

Câu 7: Liệt kê 5 gmail có số đơn hàng nhiều nhất trong tháng 10 năm 2020.

Bước 1:

Kéo thả các mục cần thiết từ bảng Fields( order\_date(year),order\_date(month),E Mail, order\_id)

Để xuất hiện mục đếm lọc trùng thì ta click chuột vào order\_id ở values, chọn Value Field settings> chọn distinct count

A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated

Sau đó ở phần filter Year và fliter Month chọn 2020 và Oct



Bước 2 : Chọn Auto sort



Chọn Value filters > chọn mục top10 và chỉnh sửa theo hình bên dưới

A screenshot of a computer

Description automatically generated

Kết quả cuối cùng sẽ giống như hình sau:

A screenshot of a spreadsheet

Description automatically generated with medium confidence

Câu 8: Hãy cho biết 4 bang có tổng số đơn hàng thấp nhất trong năm 2020

Bước 1:

Kéo thả các mục cần thiết từ bảng Fields( order\_date(year),State, order\_id)

Để xuất hiện mục đếm lọc trùng thì ta click chuột vào order\_id ở values, chọn Value Field settings> chọn distinct count

A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated

Ở phần filter Year chọn 2020



Bước 2 : Chọn Auto sort



Chọn Value filters > chọn mục top10 và chỉnh sửa theo hình bên dưới

A screenshot of a computer

Description automatically generated

Kết quả cuối cùng sẽ giống như hình sau:

A screenshot of a computer

Description automatically generated with medium confidence

Câu 9: Hãy cho biết tổng chi phí của các đơn hàng ở thành phố Alto theo từng tháng, từng năm

Bước 1:

Kéo thả các mục cần thiết từ bảng Fields( order\_date(year),City, value)

A screenshot of a computer

Description automatically generated

Sau đó ở phần city sort chọn Alto



Kết quả cuối cùng sẽ giống như hình sau:

A screenshot of a computer

Description automatically generated with low confidence

Câu 10: Hiển thị số lượng đặt hàng của độ tuổi 70 trong mỗi tháng của năm 2020

Bước 1:

Kéo thả các mục cần thiết từ bảng Fields( order\_date(year),order\_date(month),age, qty\_ordered)

A screenshot of a computer

Description automatically generated

Sau đó ở phần age fliter chọn 70



Kết quả cuối cùng sẽ giống như hình sau:

A screenshot of a computer

Description automatically generated with low confidence

# MDX

//1. Đếm số lượng ma đơn hang (order\_id) với phương thức thanh toán thành phố brownstown theo từng năm

SELECT NON EMPTY [Measures].[FACT Order Count] ON COLUMNS,

NON EMPTY([DIM Order Date].[Year].Children) ON ROWS

FROM [AMAZON SL]

WHERE [DIM Customers].[City].&[Brownstown];

A screenshot of a computer

Description automatically generated with medium confidence

//2. Cho biết khu vực (Region) có số ma đơn hàng (order\_id) nhiều nhất và số đơn hàng ít nhất

SELECT NON EMPTY [Measures].[FACT Order Count] ON COLUMNS,

NON EMPTY UNION(TOPCOUNT({[DIM Customers].[Region].Children},1,[Measures].[FACT Order Count]),

BOTTOMCOUNT({[DIM Customers].[Region].Children},2,[Measures].[FACT Order Count])) ON ROWS

FROM [AMAZON SL];

A picture containing text, screenshot, font, number

Description automatically generated

//3. Cho biết 5 thành phố có nhiều mã đơn hàng nhất trong năm 2021

SELECT NON EMPTY [Measures].[FACT Order Count] ON COLUMNS,

TOPCOUNT([DIM Customers].[City].Children,5,[Measures].[FACT Order Count]) ON ROWS

FROM [AMAZON SL]

WHERE [DIM Order Date].[Year].&[2021];

A screenshot of a computer

Description automatically generated with low confidence

//4. Hãy cho biết 3 thành phố có số lượng đơn hàng thấp nhất và tổng chi phí đơn hàng của thành phố trong năm 2020.

SELECT {[Measures].[Qty Ordered],[Measures].[Value]} on Columns,

NON EMPTY BOTTOMCOUNT(

{[DIM Customers].[City].children}

,4,[Measures].[Qty Ordered]

) ON ROWS

FROM [AMAZON SL]

WHERE [DIM Order Date].[Year].&[2020];

A picture containing text, screenshot, font, number

Description automatically generated

//5. Với từng loại sản phẩm liệt kê doanh thu có bang là Ri

SELECT NON EMPTY {[Measures].[Value]} ON COLUMNS,

NON EMPTY{[DIM Goods].[Category].children\*[DIM Order Date].[Month].children}ON ROWS

FROM [AMAZON SL]

WHERE [DIM Customers].[State].&[Ri];

A picture containing text, screenshot, font, number

Description automatically generated

//6. Chon ra 10 đơn hàng có giá trị lớn nhất tại hạt Accomack cua tung thang trong nam 2020

SELECT NON EMPTY [Measures].[Value] ON COLUMNS,

NON EMPTY TopCount ([DIM Customers].[County].[Accomack]\*[DIM Order Date].[Month].Children,10,[Measures].[Value]) ON ROWS

FROM [AMAZON SL]

WHERE [DIM Order Date].[Year].&[2020];

A screenshot of a cellphone

Description automatically generated with low confidence

//7. Hiển thị các email có số đơn hàng nhiều nhất nhất trong tháng 10 năm 2020.

SELECT NON EMPTY {[Measures].[FACT Order Count]} ON COLUMNS,

NON EMPTY TOPCOUNT({[DIM Customers].[E Mail].Children},5,[Measures].[FACT Order Count]) ON ROWS

FROM [AMAZON SL]

WHERE [DIM Order Date].[Y\_M].&[2020].&[10];

A screenshot of a computer

Description automatically generated with low confidence

//8. Hãy cho biết 4 bang có tổng số đơn hàng thấp hàng nhất trong năm 2020

SELECT NON EMPTY [Measures].[FACT Order Count] ON COLUMNS,

NON EMPTY BOTTOMCOUNT([DIM Customers].[State].Children,5,[Measures].[FACT Order Count]) ON ROWS

FROM [AMAZON SL]

WHERE [DIM Order Date].[Year].&[2020];

A picture containing text, screenshot, font, number

Description automatically generated

//9.Hãy cho biết tổng chi phí đơn hàng lớn nhất của thành phố Alto theo từng tháng ,từng năm

Select Non empty{[Measures].[Value]} on Columns,

NON EMPTY{DrillDownLevel(

DrillDownLevel([DIM Order Date].[Y\_M]))}on Rows

From [AMAZON SL]

WHERE [DIM Customers].[City].&[Alto];

A screenshot of a computer

Description automatically generated with low confidence

//10. Hiển thị số lượng đặt hàng của độ tuổi 70 trong mỗi tháng của năm 2020.

SELECT NON EMPTY [Measures].[Qty Ordered] ON COLUMNS,

NON EMPTY([DIM Order Date].[Month].Children \* [DIM Customers].[Age].&[70])ON ROWS

FROM [AMAZON SL]

WHERE [DIM Order Date].[Year].&[2020];



//Pivot

SELECT NON EMPTY

([DIM Order Date].[Month].Children \* [DIM Customers].[Age].&[70])ON COLUMNS,[Measures].[Qty Ordered] ON ROWS

FROM [AMAZON SL]

WHERE[DIM Order Date].[Year].&[2020];

A picture containing text, screenshot, font, line

Description automatically generated

# 8. DATAMINING[3]

## 8.1 Linear Regression

Khai báo các thư viện cần thiết cho quá trình datamining

A screenshot of a computer program

Description automatically generated with medium confidence

Khai báo dataset

A picture containing text, font, screenshot, clock

Description automatically generated

Chọn và in ra màn hình cột month và year

A screenshot of a computer

Description automatically generated with low confidence

In ra thông tin dataset

A screenshot of a computer program

Description automatically generated with medium confidence

Cho biết các cột của dataset

A screenshot of a computer program

Description automatically generated with medium confidence

Tạo ra bản sao của df để sử dụng trong quá trình data mining

A picture containing text, font, clock, screenshot

Description automatically generated

Bỏ các cột dưới khỏi dataset bản sao

A picture containing text, screenshot, font

Description automatically generated

Các bước chuẩn bị cần để thuật toán LR hoạt động tốt

A screen shot of a computer code

Description automatically generated with low confidence

Kết quả của dataset sau khi thực hiện các bước trên

A screenshot of a computer

Description automatically generated with low confidence

Sử dụng thư viện matplotlib để tạo đồ thị cột cho hai biến như hình dưới

A picture containing text, screenshot, font

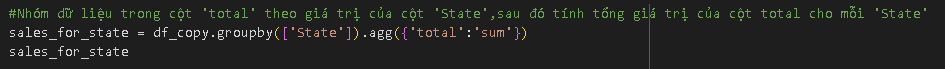
Description automatically generated

Kết quả

A picture containing screenshot, text, diagram, plot

Description automatically generated

Nhóm dữ liệu trong cột total theo giá trị của cột ‘State,sau đó tính tổng giá trị của cột total với mỗi ‘State’



Kết quả

A screenshot of a phone

Description automatically generated with low confidence

Sử dụng thư viện matplotlib để vẽ biểu đồ cho cột của biến ‘Sales of state’

A screenshot of a computer program

Description automatically generated with low confidence

Kết quả

A picture containing screenshot, plot, line, diagram

Description automatically generated

Sử dụng thư viện matplotlib và pandas để vẽ biểu đồ đường cho 2 biến sau

A screen shot of a computer code

Description automatically generated with low confidence

Kết quả

A picture containing line, plot, diagram, screenshot

Description automatically generated

Tạo một bản sao của df\_copy và lưu vào biến sales,sau đó tạo ra một danh sách của biến sale

A picture containing text, screenshot, font

Description automatically generated

Kết quả

A picture containing text, screenshot, black and white, font

Description automatically generated

In ra các giá trị mong muốn

A screen shot of a computer program

Description automatically generated with low confidence

Kết quả

A screenshot of a computer

Description automatically generated with medium confidence

Sử dụng thư viện seaborn để tạo một biểu đồ boxplot cho biến total

A screenshot of a computer

Description automatically generated with low confidence

Kết quả

A screen shot of a graph

Description automatically generated with medium confidence

Tính các giá trị cần thiết và vẽ biểu đồ

A picture containing text, screenshot

Description automatically generated

Kết quả

A picture containing screenshot, rectangle, text, diagram

Description automatically generated

In ra các giá trị mong muốn

A picture containing text, screenshot, font

Description automatically generated

Kết quả

A screenshot of a computer screen

Description automatically generated with medium confidence

Tạo một biểu đồ quantile-quantile với các thư viện statsmodels và matplotlib để kiểm tra tính chuẩn của phân phối của cột ‘total’

A screen shot of a computer

Description automatically generated with medium confidence

Kết quả

A picture containing screenshot, text, line, plot

Description automatically generated

Vẽ biểu đồ histogram của cột total cùng với đường phân phối xác xuất mật đô của phần phối chuẩn tương ứng

A picture containing text, screenshot, font

Description automatically generated

Kết quả

A picture containing text, display, screenshot, plot

Description automatically generated

Tính ma trận tương quan



Kết quả

A screenshot of a computer

Description automatically generated with low confidence

Chia train test và kiểm tra độ phù hợp của mô hình thông qua các số liệu

A picture containing text, screenshot, font

Description automatically generated

Kết quả

A picture containing text, screenshot, font

Description automatically generated

## 8.2 Random Forest

Khai báo các thư viện cần thiết cho quá trình datamining

A screen shot of a computer

Description automatically generated with medium confidence

Đọc và hiển thị 5 dòng đầu tiên của dữ liệu

A picture containing text, font, screenshot

Description automatically generated

Kết quả

A picture containing text, screenshot, font

Description automatically generated

Vẽ biểu đồ cho các giá trị bên dưới

A picture containing text, screenshot, font

Description automatically generated

Kết quả

A picture containing screenshot, text, line, plot

Description automatically generated

Ở 2 đồ thị trên ta sẽ thấy được sự tương quan cũng như mối quan hệ giữa các thuộc tính với nhau như: Doanh thu với các Bang (States) và Doanh thu với các khu vực địa lý (Regions).

Từ đây ta cũng có thể rút ra được doanh thu của khu vực địa lý hay bang nào sẽ chiếm tỉ trọng cao nhất và giảm dần về các khu vực địa lý hay bang chiếm tỉ trọng doanh thu nhỏ nhất.

Theo bang thì ta có thể thấy bang OH (Ohio) có doanh thu cao nhất và bang RI (Rhode Island) có doanh thu thấp nhất và đối với các khu vực địa lý thì Midwest chiếm doanh thu cao nhất và Northeast chiếm doanh thu nhỏ nhất.

A screen shot of a computer program

Description automatically generated with low confidence

Đoạn code trên được sử dụng để tạo hai danh sách màu sắc (colors và colors2) dựa trên giá trị tổng của nhóm dữ liệu được nhóm lại theo cột 'state' và 'region' trong dataframe df.

Đoạn code trích xuất thông tin về các tiểu bang (states) và các khu vực (regions) từ dataframe df và tính tổng doanh số bán hàng của mỗi tiểu bang và mỗi khu vực. Sau đó, nó so sánh mỗi giá trị tổng doanh số với giá trị trung bình của tổng doanh số của tất cả các tiểu bang và các khu vực. Nếu giá trị tổng doanh số của một tiểu bang hay một khu vực lớn hơn giá trị trung bình, màu 'lightgreen' được gán cho tiểu bang hay khu vực đó đó; ngược lại, màu 'lightblue' được gán cho tiểu bang hay khu vực đó. Kết quả là danh sách màu (colors) chứa các màu tương ứng với từng tiểu bang và (colors2) chứa các màu tương ứng với từng khu vực.

Vẽ biểu đồ

A computer code on a black background

Description automatically generated with low confidence

Kết quả

A screenshot of a graph

Description automatically generated with low confidence

Cả hai biểu đồ cung cấp cái nhìn tổng quan về doanh số bán hàng theo khu vực và tiểu bang, cho phép so sánh giữa các khu vực và tiểu bang về mức độ đóng góp vào tổng doanh số bán hàng. Đường kẻ ngang màu đỏ biểu thị giá trị trung bình giúp so sánh các giá trị doanh số bán hàng với mức trung bình và có thể cho thấy sự phân bố của các khu vực và tiểu bang theo mức độ doanh số bán hàng.

A screen shot of a computer program

Description automatically generated with low confidence

Đoạn code trên được sử dụng để tạo hai danh sách màu sắc (colors3 và colors4) dựa trên giá trị trung bình của nhóm dữ liệu được nhóm lại theo cột 'state' và 'region' trong dataframe df.

Đoạn code trích xuất thông tin về các tiểu bang (states) và các khu vực (regions) từ dataframe df và tính giá trị trung bình của tổng doanh số bán hàng của mỗi tiểu bang và mỗi khu vực. Sau đó, nó so sánh từng giá trị trung bình với giá trị trung bình của tổng doanh số của tất cả các tiểu bang và khu vực.

Điều này được thực hiện bằng cách lặp qua từng giá trị trung bình và kiểm tra xem giá trị đó có lớn hơn giá trị trung bình của tổng doanh số hay không. Nếu giá trị trung bình của một tiểu bang hoặc khu vực lớn hơn giá trị trung bình của tổng doanh số, màu 'lightgreen' được gán cho tiểu bang hoặc khu vực đó. Ngược lại, nếu giá trị trung bình nhỏ hơn hoặc bằng giá trị trung bình của tổng doanh số, màu 'lightblue' được gán. Kết quả là danh sách màu (colors3) chứa các màu tương ứng với từng tiểu bang và (colors4) chứa các màu tương ứng với từng khu vực.

Vẽ biểu đồ

A picture containing text, screenshot, font

Description automatically generated

Kết quả

A picture containing plot, line, text, diagram

Description automatically generated

2 đồ thị này thể hiện giá trị trung bình của doanh số bán hàng theo khu vực và tiểu bang, và so sánh giá trị trung bình đó với giá trị trung bình của tổng doanh số của tất cả các khu vực và tiểu bang. Các đường ngang đại diện cho giá trị trung bình và màu sắc của các cột thể hiện xem giá trị trung bình của khu vực hoặc tiểu bang có cao hơn hay thấp hơn giá trị trung bình chung.

Vẽ biểu đồ

A picture containing text, screenshot, font

Description automatically generated

Kết quả

A picture containing plot, line, screenshot, diagram

Description automatically generated

Cả 2 đồ thị này hiển thị tổng doanh số bán hàng theo danh mục (category) trong hai tiểu bang Iowa và Louisiana. Các biểu đồ bar chart cho phép so sánh doanh số bán hàng giữa các danh mục khác nhau trong từng tiểu bang.

Ví dụ:

Ở tiểu bang Iowa thì ta có thể thấy được doanh thu cao nhất thuộc về các sản phẩm thuộc danh mục Mobiles and Tablets và sản phẩm thuộc danh mục Men's Fashion có doanh thu thấp nhất. Và có một số doanh mục hầu như không tạo ra doanh thu ở tiểu bang này như Kids & Baby, Soghaat, v.v

Vẽ biểu đồ

A picture containing text, screenshot, font

Description automatically generated

Kết quả

A picture containing screenshot, text, plot, diagram

Description automatically generated

Tương tự ở 2 tiểu bang này thì ta có thể thấy chỉ có 1 số ít danh mục sản phẩm là có doanh thu ở 2 tiểu bang này. Đối với Vermont là 3 danh mục: cao nhất là Men's Fashion và thấp nhất là Appliances. Đối với New Mexico là 4 danh mục: cao nhất là Mobiles and Tablets và thấp nhất là Superstore. Và hầu như các danh mục khác không tạo ra doanh thu ở 2 tiểu bang này.

Đoạn code dưới trích xuất thông tin về các khu vực (region) và tiểu bang (state) từ dataframe df. Sau đó, nó tính giá trị trung bình của doanh số bán hàng (value) cho từng cặp khu vực và tiểu bang.

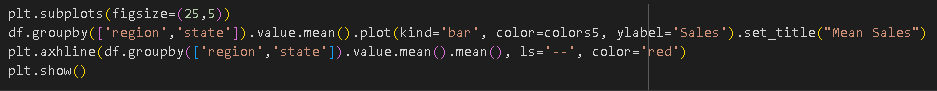
Tiếp theo, đoạn code so sánh mỗi giá trị trung bình của doanh số bán hàng với giá trị trung bình của toàn bộ doanh số bán hàng của tất cả các khu vực và tiểu bang. Nếu giá trị trung bình của doanh số bán hàng của một cặp khu vực và tiểu bang lớn hơn giá trị trung bình chung, màu 'lightgreen' được gán cho cặp khu vực và tiểu bang đó; ngược lại, màu 'lightblue' được gán.

Kết quả là danh sách màu (colors5) chứa các màu tương ứng với từng cặp khu vực và tiểu bang trong dataframe. Màu 'lightgreen' thể hiện rằng doanh số bán hàng của cặp khu vực và tiểu bang đó cao hơn giá trị trung bình chung, trong khi màu 'lightblue' thể hiện rằng doanh số bán hàng của cặp khu vực và tiểu bang đó thấp hơn giá trị trung bình chung.

A picture containing text, screenshot, font

Description automatically generated

Vẽ biểu đồ



Kết quả

A picture containing screenshot, plot, line, diagram

Description automatically generated

Biểu đồ giúp thể hiện sự khác biệt giữa doanh số bán hàng theo cặp thuộc tính khu vực và tiểu bang, và cho phép chúng ta nhận ra những tiểu bang nào ở khu vực nào có giá trị trung bình doanh số bán hàng cao hơn hoặc thấp hơn so với giá trị trung bình chung.

Nhìn vào đồ thị ta có thể thấy được các bang như IA, MN, OH, NJ, RI, LA, NC, SC, VA, CA, ID là có giá trị trung bình doanh số bán hàng cao hơn so với giá trị trung bình chung.

Vẽ biểu đồ

A picture containing text, screenshot, font

Description automatically generated

Kết quả

A picture containing text, screenshot, number, plot

Description automatically generated

Biểu đồ thứ nhất (phía trái) là một scatter plot (biểu đồ phân tán) thể hiện mối quan hệ giữa Doanh thu và các Tháng trong năm. Trục x biểu thị giá trị Doanh thu, trong khi trục y biểu thị các tháng. Mỗi điểm trên biểu đồ đại diện cho một giá trị 'value' tương ứng với tháng tương ứng. Đường thẳng màu đỏ là đường trung bình của biến 'value' và đại diện cho giá trị trung bình của 'value' trên toàn bộ dữ liệu. Biểu đồ này giúp xem xét mối quan hệ giữa 'value' và 'month' và nhận biết sự phân bố của giá trị 'value' trong từng tháng.

Ở đồ thị trái ta có thể thấy tổng doanh thu sẽ chiếm tỉ trọng cao (chỉ là chiếm nhiều hơn so với các tháng khác chứ không quá biến động) ở 2 tháng 10 và 12. Còn ở các tháng khác thì doanh số trải đều không bị trường hợp có tháng doanh thu quá nhỏ hay doanh thu bị thưa giữa các tháng.

Biểu đồ thứ hai (phía phải) cũng là một scatter plot thể hiện mối quan hệ giữa Doanh thu và số lượng của sản phẩm trong đơn hàng. Trục x biểu thị giá trị 'value', trong khi trục y biểu thị số lượng hàng đã đặt. Mỗi điểm trên biểu đồ đại diện cho một giá trị 'value' và số lượng hàng đã đặt tương ứng. Biểu đồ này giúp xem xét mối quan hệ giữa 'value' và 'qty\_ordered' và nhận biết sự phân bố của giá trị 'value' đối với số lượng hàng đã đặt.

Ở đồ thị phải ta có thể thấy tổng doanh thu sẽ tập trung ở những đơn hàng có số lượng đặt hàng của 1 sản phẩm từ 11 trở xuống và có giá trị của biến 'value' là tầm 2500 trở xuống. Và đa số các đơn hàng có số lượng đặt hàng của 1 sản phẩm là trên 12 hay 13 đều chiếm doanh thu rất ít.

Dự đoán doanh số sử dụng thuật toán Random Forest

Khai báo thư viện cần thiết

A picture containing text, screenshot, font

Description automatically generated

Đếm số lượng đơn hàng theo category

A screenshot of a computer

Description automatically generated

Vẽ biểu đồ

A picture containing screenshot, text, font

Description automatically generated

Kết quả

A picture containing text, screenshot, number, parallel

Description automatically generated

Cả hai biểu đồ đều giúp thể hiện phân phối và mối quan hệ giữa doanh số và tháng trong toàn bộ tập dữ liệu (subplot 1) và trong danh mục 'Mobiles & Tablets' (subplot 2). Điều này giúp phân tích và so sánh doanh số theo thời gian và trong một danh mục cụ thể.

Thì ở Biểu đồ phía bên phải ta có thể thấy Doanh thu của danh mục Mobiles & Tablets sẽ đạt ở mức cao nhất từ tháng 10 cho đến tháng 12 (của năm 2020)

Tạo một biến dummie và để giá trị của cột category vào , sau đó bỏ cột category

A picture containing text, font, screenshot, line

Description automatically generated

In ra 5 giá trị đầu của dataset sau khi đã làm như bước trên

A screenshot of a computer

Description automatically generated with medium confidence

Bỏ các cột sau

A screen shot of a computer

Description automatically generated with low confidence

A picture containing text, screenshot, font

Description automatically generated

Đoạn code trên thực hiện các bước sau:

Train/test split: Dữ liệu được chia thành hai tập là tập huấn luyện (x\_train, y\_train) và tập kiểm tra (x\_test, y\_test) bằng hàm train\_test\_split(). Tập kiểm tra chiếm 20% tổng số dữ liệu ban đầu. Đây là bước quan trọng để đánh giá hiệu suất của mô hình trên dữ liệu mới.

Khởi tạo mô hình: Mô hình RandomForestRegressor được khởi tạo bằng cách tạo một đối tượng forest thông qua RandomForestRegressor().

Huấn luyện mô hình: Mô hình RandomForestRegressor được huấn luyện trên dữ liệu huấn luyện (x\_train, y\_train) bằng cách gọi phương thức fit(). Mô hình học từ dữ liệu huấn luyện để tìm mối quan hệ giữa các đặc trưng (x\_train) và biến mục tiêu (y\_train).

Dự đoán: Mô hình được sử dụng để dự đoán giá trị của biến mục tiêu trên tập dữ liệu kiểm tra (x\_test) bằng cách gọi phương thức predict(). Kết quả dự đoán được lưu vào biến predicao.

Kết quả

A picture containing text, font, typography

Description automatically generated

Đánh giá mô hình: Đoạn code tính toán và in ra các độ đo đánh giá hiệu suất của mô hình trên tập kiểm tra.

A screenshot of a computer

Description automatically generated with low confidence

Kết quả



# 9.REFERENCES

[1] *Hướng dẫn tạo project SSIS SSAS đơn giản với SQL server và Visual Studio 2022*, (2023). Accessed: Jun. 30, 2023. [Online Video]. Available: https://www.youtube.com/watch?v=Vz8ttUSKpYo

[2] *Thành thạo Pivot table excel trong 17 phút*, (2020). Accessed: Jun. 30, 2023. [Online Video]. Available: https://www.youtube.com/watch?v=sqFzDYlm9H0

[3] “Kaggle: Your Machine Learning and Data Science Community.” https://www.kaggle.com/ (accessed Jun. 30, 2023).