TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN – ĐHQG HCM

KHOA CÔNG NGHỆ THÔNG TIN

MÔN: HỆ THỐNG THÔNG TIN PHỤC VỤ TRÍ TUỆ KINH DOANH 17\_11

ĐỒ ÁN MÔN HỌC LẦN 3

PHÂN TÍCH VÀ THIẾT KẾ KHO DỮ LIỆU

Nhóm: CQ-BI-05

1712055 – Ngô Kim Quốc Huy

1712072 – Nguyễn Văn Khoa

1712112 – Cao Thành Nhân  
**1712150 – Phạm Đỗ Nhật Tân (L)**

Mục lục

[1. Thông tin thành viên và phân công 2](#_Toc61302418)

[2. Xác định yêu cầu 2](#_Toc61302419)

[3. Cube 4](#_Toc61302422)

[3.1. Thiết kế cube 4](#_Toc61302426)

[3.2. Phân chiều trên các dim 5](#_Toc61302431)

[3.3. Truy vấn MDX 5](#_Toc61302437)

[4. Mining 6](#_Toc61302438)

[5. Report 7](#_Toc61302439)

# Thông tin thành viên và phân công

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| MSSV | Họ tên | Công việc | Nhóm trưởng | Đóng góp |
| 1712055 | Ngô Kim Quốc Huy | MDX: YC1  Report: YC8  Tiền xử lý dữ liệu |  | 10% |
| 1712072 | Nguyễn Văn Khoa | MDX: YC4  Report: YC10  Thu thập và xử lý dữ liệu |  | 10% |
| 1712112 | Cao Thành Nhân | MDX: YC2, YC1, YC3  Report: YC9, YC10  Lấy location bằng API, thu thập và xử lý dữ liệu. |  | 70% |
| 1712150 | Phạm Đỗ Nhật Tân | MDX: YC3.  Report: YC11  Build Cube, lấy location bằng API, điều chỉnh ETL. | x | 10% |

**Các yêu cầu phức tạp khác được thực hiện chung bởi tất cả các thành viên khác**

# Xác định yêu cầu

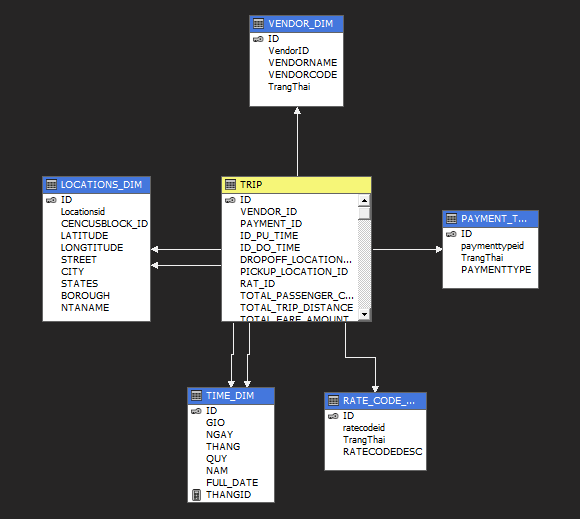


|  |  |  |  |
| --- | --- | --- | --- |
| Identifier | Name | Description | Type |
| 1 | Thống kê doanh thu, số lượng khách | *Thống kê doanh thu, số lượng khách theo tháng, quý.* | MDX |
| 2 | Phân tích doanh thu các chuyến đi | *Thống kê tổng doanh thu của các chuyến đi theo Street, Borough* | MDX |
| 3 | Thống kê lượng thuế, phí, tiền boa | *Thống kê lượng thuế, phí, tiền boa theo tháng của từng năm* | MDX |
| 4 | Thống kê số chuyến đi | *Thống kế số lượng chuyến đi theo từng borough* | MDX |
| 5 | Thống kê số lượng đón taxi theo giờ | *Phân tích thời điểm nào trong ngày là giờ cao điểm và thấp điểm của*  *các chuyến taxi. Vẽ đồ thị phân bố số lượng đón taxi theo giờ trong*  *ngày trong tất cả các năm.* | Vẽ biểu đồ |
| 6 | Dùng regional map biểu diễn doanh thu | *Dùng regional map biểu diễn doanh thu theo borough NTANAME STREET* | Vẽ biểu đồ |
| 7 | Phân bố lượt đón theo CENSUS BLOCK | *Dùng regional map phân bố lượt đón theo CENSUS BLOCK* | Vẽ biểu đồ |
| 8 | Phân bố chuyến đi theo dòng thời gian | *Vẽ được biểu đồ phân bô số chuyến đi theo thời gian* | Vẽ biểu đồ |
| 9 | Phân bố chuyến đi theo doanh thu và tips | *Vẽ được biểu đồ phân bô số chuyến đi theo* doanh thu và tips theo dòng thời gian | Vẽ biểu đồ |
| 10 | Biểu đồ doanh thu theo thành phố | *Vẽ biểu đồ thống kê doanh thu theo thành phố* | Vẽ biểu đồ |
| 11 | Thống kê lượt đón theo borough | *Vẽ biểu đồ thống kê doanh thu theo borough* | Vẽ biểu đồ |
| 12 | Dự đoán lượng khách đón taxi trong năm | *Dự đoán vào một thời điểm cụ thể trong năm theo từng district thì*  *lượng khách đón taxi sẽ như thế nào* | Mining |
| 13 | Thưc hiện dự đoán theo khu vực | *Để phân chia thành phố new york thành khu vực để có thể thực hiện dự đoán theo khu vực -> sử dụng thuật toán K-mean...* | Mining |
| 14 | Xây dựng mô hình dự đoán tiền boa/tổng giá vé | *Xây dựng một mô hình dự đoán cho số tiền boa theo tỷ lệ phần trăm*  *của tổng giá vé* | Mining |

# Cube



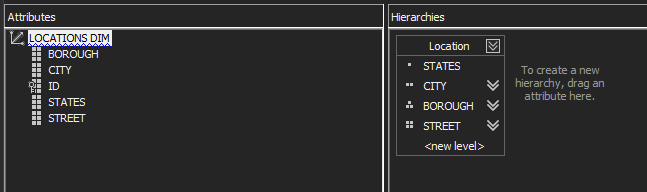
## Thiết kế cube





## Phân chiều trên các dim

Location

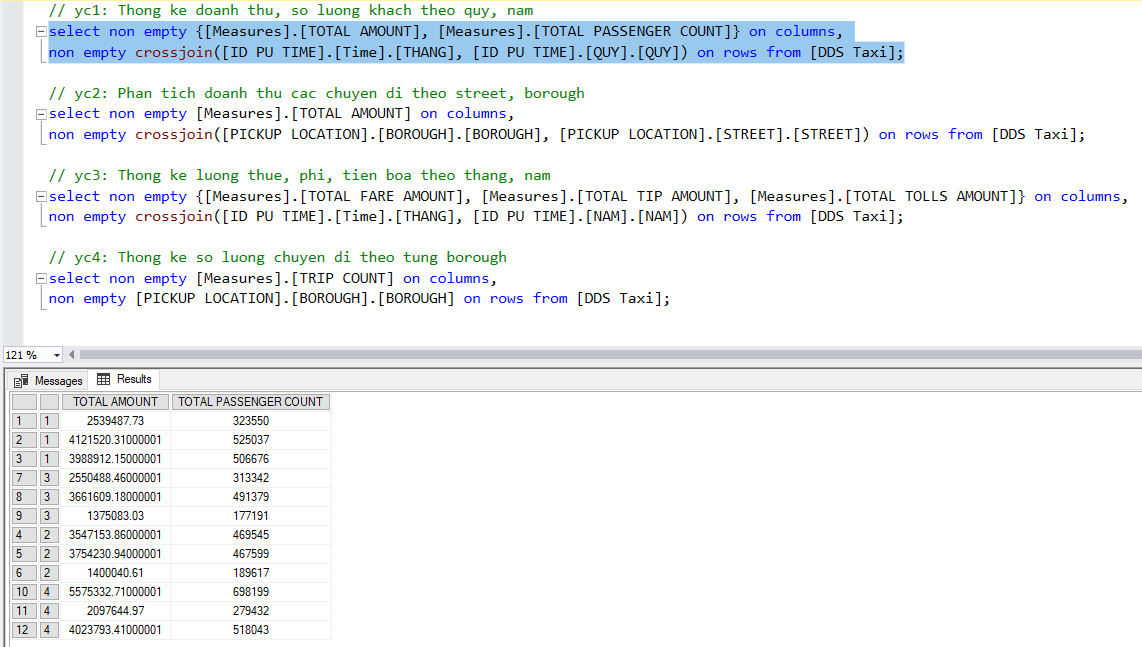


Time

\



## Truy vấn MDX



# Mining

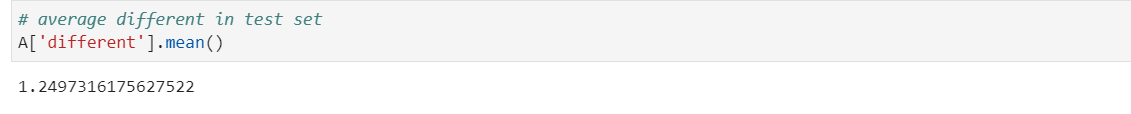
Thực hiện dự doán sử dụng Python

**Bước 1**: Sử dụng thuật toán Kmean để phân lớp dữ liệu

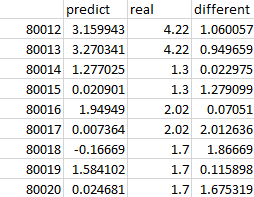
* Lấy giá trị kinh độ, vĩ độ để phản hồi
* Sử dụng độ do WCSS để xác định số cluster tốt nhất với bài toán
* Thực hiện Kmean với số lượng cluster đã chọn ở trên

**Bước 2**: Sử dụng linear regression dể dự đoán tiền boa

* Chia data theo tỉ lệ 80:20 để training và testing
* Xử lý những trường dữ liệu non-numberic
* Xây dựng model
* So sánh kết quả mình dự đoán với giá trị thực tế



**Bước 3**: Xuất kết quả ra file result.csv



# Report

