A picture containing logo

Description automatically generated

ĐẠI HỌC QUỐC GIA TPHCM – TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

KHOA CÔNG NGHỆ THÔNG TIN – BỘ MÔN HỆ THỐNG THÔNG TIN

**ĐỒ ÁN THỰC HÀNH**

**Môn học: Hệ thống thông tin phục vụ Trí tuệ Kinh doanh**

**GVHD**:

Hồ Thị Hoàng Vy

Tiết Gia Hồng

Nguyễn Thị Như Anh

**Nhóm**: TTKD-03

18120213 – Võ Đại Nam

18120214 – Lê Ngọc Bảo Ngân

18120247 – Phạm Hồ Ngọc Trâm

18120456 – Lại Bùi Thành Luân

*Thành phố Hồ Chí Minh, tháng 11 năm 2021*

**THÔNG TIN CHI TIẾT NHÓM**

|  |  |  |
| --- | --- | --- |
| **Tên nhóm:** | TTKD-03 | |
| **Số lượng:** | 04 | |
| **MSSV** | **Họ tên** | **Email** |
| 18120213 | Võ Đại Nam | [18120213@student.hcmus.edu.vn](mailto:18120213@student.hcmus.edu.vn) |
| 18120214 | Lê Ngọc Bảo Ngân | [18120214@student.hcmus.edu.vn](mailto:18120214@student.hcmus.edu.vn) |
| 18120247 | Phạm Hồ Ngọc Trâm | [18120247@student.hcmus.edu.vn](mailto:18120247@student.hcmus.edu.vn) |
| 18120456 | Lại Bùi Thành Luân | [18120456@student.hcmus.edu.vn](mailto:18120456@student.hcmus.edu.vn) |

**PHÂN CÔNG & ĐÁNH GIÁ HOÀN THÀNH CÔNG VIỆC**

|  |  |  |  |
| --- | --- | --- | --- |
| **Người**  **thực hiện** | **Công việc thực hiện** | **Mức độ hoàn thành** | **Đánh giá của nhóm** |
| 18120213  Võ Đại Nam | Quy trình ETL từ Source vào Stage của nguồn dữ liệu UK Car Accidents | 100% | 10/10 |
| 18120214  Lê Ngọc Bảo Ngân | Tổng hợp và viết báo cáo | 100% | 10/10 |
| 18120247  Phạm Hồ Ngọc Trâm | Quy trình ETL từ Source vào Stage của nguồn dữ liệu UK – Postcodes | 100% | 10/10 |
| 18120456  Lại Bùi Thành Luân | Quy trình ETL từ Source vào Stage của nguồn dữ liệu LSOA – Postcode mapping | 100% | 10/10 |

# Mục lục

[Mục lục 3](#_Toc92657465)

[Describe data 4](#_Toc92657466)

[I. UK Car Accidents 4](#_Toc92657467)

[1. Accidents 4](#_Toc92657468)

[2. Casualities 7](#_Toc92657469)

[3. Vehicles 9](#_Toc92657470)

[II. LSOA – Postcode mapping 12](#_Toc92657471)

[III. UK – Postcodes 15](#_Toc92657472)

[ETL process 17](#_Toc92657473)

[I. Source to Stage 17](#_Toc92657474)

[II. Stage to NDS 17](#_Toc92657475)

[III. NDS to DDS 18](#_Toc92657476)

[TRỰC QUAN HÓA DỮ LIỆU 20](#_Toc92657477)

[OLAP 31](#_Toc92657478)

[Link Github 47](#_Toc92657479)

[Link Drive nhóm 47](#_Toc92657480)

# 

# Describe data

## UK Car Accidents

### Accidents

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Thuộc tính** | **Mô tả** | **Kiểu Dữ liệu** | **Kiểu Dữ liệu MS SQL** | **Kiểu dữ liệu Null (%)** |
| Accident\_Index | Xác định duy nhất mã của vụ tai nạn | Chuỗi kí tự không dấu độ dài 13 | NVARCHAR (13) | 0 |
| Accident\_Severity | Mức độ nghiêm trọng của tai nạn  (1 - Fatal; 2 - Serious; 3 - Slight) | Dữ liệu dạng số | INT | 0 |
| Number\_of\_Vehicles | Số lượng xe cô trong vụ tai nạn | Dữ liệu dạng số | INT | 0 |
| Number\_of\_Casualties | Số lượng người bị thương | Dữ liệu dạng số | INT | 0 |
| Date | Ngày diễn ra vụ tai nạn | Dữ liệu ngày, tháng, năm | DATETIME | 0 |
| Time | Giờ diễn ra tai nạn | Dữ liệu giờ, phút, giây | TIME (7) | 0 |
| Local\_Authority\_(District) | Quận của địa phương diễn ra tai nạn (mỗi quận sẽ được đánh dấu số từ 1 tới 941) | Dữ liệu dạng số | INT | 0 |
| Local\_Authority\_(Highway) | Mã của địa phương diễn ra tai nạn (Mã của mỗi quận sẽ có kí tự đầu thuộc 1 trong 3 chữ cái (E, W, S) và phần còn lại là chuỗi số) | Chuỗi kí tự không dấu độ dài 9 | NVARCHAR (9) | 0 |
| Road\_Type | Loại đường ở nơi diễn ra tai nạn (mỗi loại đường được gắn 1 số trong dãy số: 1, 2, 3, 6, 7, 9, 13. Loại đường với giá trị -1 được tính là dữ liệu bị mất) | Dữ liệu dạng số | INT | 0 |
| Speed\_Limit | Xác định tốc độ tối đa | Dữ liệu dạng số | INT | 0 |
| Urban\_Or\_Rural\_Area | Xác định khu vực diễn ra tai nạn là Urban hay Rural | Dữ liệu dạng số | INT | 0 |
| LSOA\_of\_Accident\_Location | Xác định mã của LSOA tại khu vực tai nạn | Chuỗi kí tự không dấu độ dài 9 | NVARCHAR (9) | 7 |

#### Data Profiling

1. **Null value**

Table

Description automatically generated

1. **Distinct value**

Table

Description automatically generated

### Casualities

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Thuộc tính** | **Mô tả** | **Kiểu Dữ liệu** | **Kiểu Dữ liệu MS SQL** | **Kiểu dữ liệu Null (%)** |
| Accident\_Index | Xác định duy nhất mã của vụ tai nạn | Chuỗi kí tự không dấu độ dài 13 | NVARCHAR (13) | 0 |
| Casualty\_Reference | Xác định mã tham khảo của nạn nhân | Dữ liệu dạng số | INT | 0 |
| Age\_of\_Casualty | Tuổi của nạn nhân | Dữ liệu dạng số | INT | 0 |
| Vehicel\_Reference | Xác định loại tham khảo của xe cộ | Dữ liệu dạng số | INT | 0 |
| Sex\_of\_Casualty | Xác định giới tính của nạn nhân (1 - Nam; 2 - Nữ; 3 - Không rõ; -1 - dữ liệu bị mất) | Dữ liệu dạng số | INT | 0 |
| Age\_Band\_of\_Casualty | Phạm vi độ tuổi của nạn nhân trong tai nạn (mỗi phạm vi sẽ được gắn giá trị từ 1 tới 11) | Dữ liệu dạng số | INT | 0 |
| Casualty\_Severity | Xác định mức độ chấn thương của nạn nhân (1 - Fatal; 2 - Serious; 3 - Slight) | Dữ liệu dạng số | INT | 0 |
| Casualty\_Type | Xác định loại nạn nhân (mỗi loại được gắn số từ 1 tới 23, có các loại sẽ được gắn 1 trong 3 số sau: 90, 97, 98) | Dữ liệu dạng số | INT | 0 |

#### Data Profiling

1. **Null value**

Table

Description automatically generated

1. **Distinct value**

Table

Description automatically generated

### Vehicles

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Thuộc tính** | **Mô tả** | **Kiểu Dữ liệu** | **Kiểu Dữ liệu MS SQL** | **Kiểu dữ liệu Null (%)** |
| Accident\_Index | Xác định duy nhất mã của vụ tai nạn | Chuỗi kí tự không dấu độ dài 13 | NVARCHAR (13) | 0 |
| Vehicle\_Type | Xác định loại xe cộ (mỗi loại được gắn số từ 1 tới 23, có các loại sẽ được gắn 1 trong 3 số sau: 90, 97, 98) | Dữ liệu dạng số | INT | 0 |
| Vehicel\_reference | Xác định loại tham khảo của xe cộ | Dữ liệu dạng số | INT | 0 |
| Journey\_Purpose\_of\_Driver | Xác định mục tiêu di chuyển của tài xế | Dữ liệu dạng số | INT | 0 |
| Sex\_of\_Driver | Xác định giới tính của tài xế (1 - Nam; 2 - Nữ; 3 - Không rõ; -1 - dữ liệu bị mất) | Dữ liệu dạng số | INT | 0 |
| Age\_of\_Driver | Tuổi của tài xế | Dữ liệu dạng số | INT | 0 |
| Age\_Band\_of\_Driver | Phạm vi độ tuổi của tài xế (mỗi phạm vi sẽ được gắn giá trị từ 1 tới 11) | Dữ liệu dạng số | INT | 0 |
| Engine\_Capacity\_(CC) | Công suất khối của động cơ | Dữ liệu dạng số | INT | 0 |

#### Data Profiling

1. **Null value**

Table

Description automatically generated

1. **Distinct value**

Table

Description automatically generated

## LSOA – Postcode mapping

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Thuộc tính** | **Mô tả** | **Phạm vị giá trị** | **Nhận xét** | **Số dòng null** |
| pcd7 | Mã bưu điện với 7 ký tự | AB1Δ1AA 🡪 ZE999ZZ | Tất cả các mã bưu điện hiện tại ("live") trong Vương quốc Anh, Quần đảo Channel và Đảo Man, nhận được hàng tháng từ Royal Mail.  Ngoài ra, tất cả các mã bưu chính đã chấm dứt (‘closed’) sau đó không được sử dụng lại bởi Royal Mail trong Vương quốc Anh và các cơ quan bưu chính ở Quần đảo Channel và Đảo Man.  2, 3 hoặc 4 ký tự mã hướng ra ngoài - căn trái;  3 ký tự mã hướng nội - căn phải;  Ký tự thứ 3 và thứ 4 có thể để trống (Δ) | 0 |
| pcd8 | Mã bưu điện với 8 ký tự | ABΔΔ1AA 🡪 ZE9Δ9ZZ | Như trên, ngoại trừ:  Mã hướng ra ngoài 2, 3 hoặc 4 ký tự - căn trái;  3 ký tự hướng nội mã - căn phải;  Ký tự thứ 5 luôn trống và có thể ký tự thứ 3 và thứ 4 cũng để trống (Δ) | 0 |
| pcds | Mã bưu điện với dạng e-Gif | AB1 1AA 🡪 ZE99 9ZZ | Mã hướng ra 2, 3 hoặc 4 ký tự Dấu cách đơn (Δ)  Mã hướng đi 3 ký tự | 0 |
| dointr | Ngày giới thiệu | YYYYMM | Lần xuất hiện gần đây nhất của ngày giới thiệu mã bưu điện. | 0 |
| doterm | Ngày kết thúc | YYYYMM hoặc null | Nếu có, là lần xuất hiện gần đây nhất của ngày chấm dứt mã bưu điện, nếu không: null = ‘live’ mã bưu điện | 68436 |
| usertype | Loại người dùng mã bưu điện | 0 hoặc 1 | Cho biết mã bưu điện là của người dùng nhỏ hay lớn.  0 = người dùng nhỏ;  1= người dùng lớn; | 0 |
| oa11cd | Khu vực đầu ra tương ứng với mã bưu điện | none | Dùng để tìm kiếm toạ độ tương ứng với mã | 200 |
| Isoa11cd | Khu vực đầu vào tương ứng với mã bưu điệnn | None | Dùng để tìm kiếm toạ độ đầu vào tương ứng với mã | 200 |
| msoa11cd | Khu vực đầu ra ở giữa | None | Dùng để tìm kiếm tầng giữa của các siêu khu vực | 200 |
| ladcd | Mã chính quyền địa phương hiện tại các quận | None | Dùng để tìm kiếm chính quyền địa phương hiện tại các quận | 200 |
| lsoa11nm | Tên khu vực đầu vào | None | Tên viết bằng chữ của số nhà đường chi tiết của khu vực đầu vào tại Anh | 200 |
| msoa11nm | Tên khu vực trung tâm | None | Tên viết bằng chữ của số nhà, đường chi tiết của khu vực trung tâm vào tại Anh | 200 |
| ladnm | Tên quận, thành phố được đề cập | None | Tên viết bằng chữ của Quận, Thành phố trực thuộc của khu vực được đề cập tại Anh | 200 |
| ladnmw | Tên quận, thành phố được đề cập mới thay đổi | None | Tên viết bằng chữ của Quận, Thành phố trực thuộc của khu vực được đề cập tại Anh mới thay đổi | 100% |

#### Data Profiling

1. **Null value**

Chart

Description automatically generated with medium confidence

1. **Distinct value**

Ảnh có chứa bàn

Mô tả được tạo tự động

## UK – Postcodes

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Thuộc tính** | **Mô tả** | **Kiểu dữ liệu** | **Ràng buộc** | **Kiểu dữ liệu**  **MS SQL** |
| Easting | Tọa độ x (tính về hướng Đông) | Số thực | PK | VARCHAR (50) |
| Northing | Tọa độ y (tính về hướng Bắc) | Số thực | Chưa biết | VARCHAR (50) |
| Latitude | Vĩ độ | Số thực | Chưa biết | VARCHAR (50) |
| Longitude | Kinh độ | Số thực | Chưa biết | VARCHAR (50) |
| City | Thành phố | Chuỗi kỹ tự | Chưa biết | VARCHAR (50) |
| County | Hạt | Chuỗi kỹ tự | Chưa biết | VARCHAR (50) |
| Country\_code | Mã số quốc gia | Chuỗi kỹ tự | Chưa biết | VARCHAR (50) |
| Country\_name | Tên quốc gia | Chuỗi kỹ tự | Chưa biết | VARCHAR (50) |
| Iso3166-2 | Mã ISO của quốc gia | Chuỗi kỹ tự | Chưa biết | VARCHAR (50) |
| Region\_code | Mã số vùng miền | Chuỗi kỹ tự | Chưa biết | VARCHAR (50) |
| Region\_name | Tên vùng miền | Chuỗi kỹ tự | Chưa biết | VARCHAR (50) |

#### Data Profiling

1. **Null value**

Ảnh có chứa bàn

Mô tả được tạo tự động

1. **Distinct value**

Graphical user interface

Description automatically generated with low confidence

# ETL process

## Source to Stage

**B1**: Đổ dữ liệu thô từ Source vào trong Stage

* Tạo Data Flow Task
* Ở trong Data Flow, ta thực hiện các thao tác:
* Tạo OLE DB Source để lấy dữ liệu thô từ Source
* Tạo Lookup để kiểm tra nếu dữ liệu này đã có tồn tại trong Stage hay chưa. Nếu có tồn tại, ta cập nhật lại thời gian thêm dữ liệu vào Stage; và ngược lại, ta đặt giá trị thời gian thêm dữ liệu vào Stage là thời điểm hiện tại và sau đó tiến hành thêm mới dữ liệu này vào.
* Tạo Union All để kết hợp 2 luồng điều kiện trên
* Tạo OLE DB Destination để đổ dữ liệu sau khi đã được rút trích vào trong Stage

**B2**: Lấy giá trị LSET trong Metadata (thời điểm mới nhất khi đổ dữ liệu từ Source vào trong Stage)

* Tạo Execute SQL Task
* Mở Execute SQL Task Editor 🡪 chọn tab Result Set 🡪 Tạo LSET variable
* Nối Execute SQL Task này vào trong Data Flow Task mới tạo ở **B1** để lấy giá trị LSET trong Stage

**B3**: Cập nhật lại giá trị LSET trong Metadata

* Tạo Execute SQL Task
* Mở Execute SQL Task Editor 🡪 Chọn tab General 🡪 nhập câu query update giá trị LSET trong Metadata ở ô SQLStatement
* Nối Data Flow Task tạo ở **B1** vào Execute SQL Task này để thực hiện việc cập nhật lại giá trị LSET

## Stage to NDS

Tương tự như quá trình đổ dữ liệu từ Source vào trong Stage

**B1**: Đổ dữ liệu từ Stage vào NDS

* Tạo Data Flow Task
* Ở trong Data Flow, ta thực hiện các thao tác:
* Tạo OLE DB Source để lấy dữ liệu từ Stage
* Tạo Lookup để kiểm tra nếu dữ liệu này đã có tồn tại trong NDS hay chưa. Nếu có tồn tại, ta cập nhật lại thời gian thêm dữ liệu vào NDS; và ngược lại, ta đặt giá trị thời gian thêm dữ liệu vào NDS là thời điểm hiện tại và sau đó tiến hành thêm mới dữ liệu này vào.
* Tạo Union All để kết hợp 2 luồng điều kiện trên
* Tạo OLE DB Destination để đổ dữ liệu sau khi đã được rút trích vào trong NDS

Đối với bảng LSOA\_Stage, ta thêm vào 1 bước sử dụng Derived Column để phân tách cột pcd7 (trước bước Lookup) dùng cho mục đích thiết kế bảng Fact ở bước tiếp theo (NDS to DDS)

## NDS to DDS

**B1**: Tiến hành phân tích dữ liệu và xác định các bảng Dim: gồm 5 bảng

* ***Dim\_Geography***: chiều địa lý. Gồm các thuộc tính:
* City
* County
* County\_name
* Region\_name
* Thêm mới: GeographyID (surrogate key)

Thực hiện:

* Tạo OLE DB Source để lấy ra các giá trị từ các thuộc tính City, County, County\_name, Region\_name từ bảng Postcode\_NDS
* Thêm mới thuộc tính GeographyID làm surrogate key
* Tạo OLE DB Destination để lưu dữ liệu vào bảng Dim\_Geography
* ***Dim\_Date***: chiều thời gian. Gồm các thuộc tính:
* Day
* Month
* Year
* Quarter
* DateKey

Thực hiện:

* Tạo OLE DB Source để lấy ra dữ liệu thời gian tai nạn trong bảng Accidents\_NDS
* Sử dụng Script Component để transform dữ liệu thời gian trên thành các thuộc tính: Year, Month, Date, Quarter
* Tạo OLE DB Destination để lưu dữ liệu vào bảng Dim\_Date
* ***Dim\_Casualties***: chiều thương vong. Gồm các thuộc tính:
* AgeBandOfCasualty
* AgeOfCasualty
* CasualtySeverity
* CasualtyType
* SexofCasualty
* Thêm mới: AgeBandName

Thực hiện:

* Tạo OLE DB Source để lấy ra các dữ liệu trong bảng Casualties\_NDS
* Sử dụng Script Component để thêm thuộc tính mới AgeBandName
* Sử dụng Lookup để tra cứu các code của các thuộc tính SexOfCasualty, AgeBandOfCasualty, CasualtySeverity, CasualtyType ở trong database Codebook
* Sử dụng Slowly Changing Dimension để xác định chiều thay đổi chậm: chuyển Change Type của tất cả các cột thành giá trị Fixed Attribute
* Tạo OLE DB Destination để lưu dữ liệu vào bảng Dim\_Casualties
* ***Dim\_Vehicles***: chiều thời gian. Gồm các thuộc tính:
* JourneyPurpose
* Status
* VehicleType
* Thêm mới: VehicleID (surrogate key)

Thực hiện:

* Tạo OLE DB Source để lấy ra các dữ liệu trong bảng Vehicles\_NDS
* Sử dụng Lookup để tra cứu các code của các thuộc tính JourneyPurpose và VehicleType ở trong database Codebook
* Sử dụng Slowly Changing Dimension để xác định chiều thay đổi chậm: chuyển Change Type của các thuộc tính JourneyPurpose và VehicleType thành giá trị Historical Attribute
* Tạo OLE DB Destination để lưu dữ liệu vào bảng Dim\_Vehicles
* ***Dim\_Accidents***: chiều tai nạn. Gồm các thuộc tính:
* AccidentSeverity
* LocalAuthorityDistrict
* UrbanOrRural
* RoadType
* Thêm mới: Built-upRoad
* Thêm mới: AccidentID (surrogate key)

Thực hiện:

* Tạo OLE DB Source để lấy ra các dữ liệu trong bảng Accidents\_NDS
* Sử dụng Lookup để tra cứu các code của các thuộc tính AccidentSeverity, LocalAuthorityDistrict, UrbanOrRural, RoadType ở trong database Codebook
* Sử dụng Slowly Changing Dimension để xác định chiều thay đổi chậm: chuyển Change Type của các thuộc tính AccidentSeverity, LocalAuthorityDistrict, UrbanOrRural, RoadType và Built-upRoad thành giá trị Fixed Attribute
* Tạo OLE DB Destination để lưu dữ liệu vào bảng Dim\_Accidents

**B2**: Xây dựng bảng Fact gồm các thuộc tính:

* AccidentID
* VehicleID
* CasualtyID
* DateKey
* GeographyID
* Thuộc tính measure: Number of Vehicles, Number of Casualties

Thực hiện:

* Merge các bảng trong NDS theo key
* Lookup các bảng trong Codebook để thay đổi datatype từ code sang string
* Merge tất cả thành 1 bảng thống nhất
* Lookup các bảng Dim và lưu lại vào bảng Fact (OLE DB Destination)

# TRỰC QUAN HÓA DỮ LIỆU

Câu 1:

Top 10 địa phương có số ca tai nạn nhiều nhất

Chart, bubble chart

Description automatically generated

Nhận xét: Dựa vào mức độ nghiêm trọng, ta thấy được hầu hết số ca tai nạn là ở mức độ nhẹ, ca tử vong không nhiều, thành phố có nhiều nạn nhân bị tai nạn nhất là Birmingham.

Câu 2:

Top 10 địa phương có số ca tai nạn nhiều nhất

Treemap chart

Description automatically generated with medium confidence

Nhận xét: Biểu đồ tree map thể hiện số lượng nạn nhân bị tai nạn ở quý 3 và quý 4 khá nhiều. Dựa vào số liệu chi tiết thì số nạn nhân của quý 1 và quý 2 gần giống nhau.

Câu 3:

Top 10 loại nạn nhân có nhiều ca tai nạn nhất

Chart, bubble chart

Description automatically generated

Nhận xét: Số lượng người bị tử vong là nam hay nữ không có sự chênh lệch nhiều. Người tử vong nhiều nhất của cả 2 giới là người đi xe hơi.

Câu 4:

Chart, bar chart

Description automatically generated

Nhận xét: Số lượng tai nạn có mức độ nhẹ là nhiều nhất, tổng số ca tai nạn vào buổi chiều là nhiều nhất và vào buổi đêm là ít nhất.

Câu 5:

Chart, treemap chart

Description automatically generated

Nhận xét: Số lượng ca tai nạn chủ yếu xảy ra ở vùng thành thị, tuy nhiên đa số là ca nhẹ.

Câu 6:

Chart, bubble chart

Description automatically generated

Nhận xét: Đa số ca tai nạn là ca nhẹ với loại nạn nhân là người lớn lái ô tô.

Câu 7:

A picture containing graphical user interface

Description automatically generated

Nhận xét: Từ sự chênh lệch khá lớn của biểu đồ có thể thấy được người bị tai nạn chủ yếu là đi xe hơi và bị tai nạn trên lúc đi làm hoặc lý do khác vì số lượng lý do chưa biết cũng là khá lớn.

Câu 8:

Chart, treemap chart

Description automatically generated

Nhận xét: Số lượng ca tai nạn diễn ra khá nhiều trong đường built-up road với loại phương tiện là xe hơi dù tốc độ giới hạn của đường chỉ dưới 50mph.

Câu 13.

Visualize Map:

Khu vực diễn ra tai nạn theo country:

Map

Description automatically generated

Khu Vực diễn ra tai nạn theo Region:

Map

Description automatically generated

Khu vực diễn ra tại nạn theo county:

Map

Description automatically generated

Khu vực diễn ra tai nạn theo city:

Map

Description automatically generated

Case Study: Drill Though Scotland để thấy sự phân bố tại nạn trong các năm

Drill Though County Của Scotland

Map

Description automatically generated

Drill though xuống city:

A picture containing diagram

Description automatically generated

# OLAP

Mô hình Cube:

Diagram

Description automatically generated

Câu 1: Thống kê số lượng nạn nhân theo Mức Độ Nghiêm Trọng (Fatal, Serious,

Slight) ở các Địa phương (Local\_Authority\_(District)) trong tất cả các năm.

OLAP:

Measure: Number or casualties

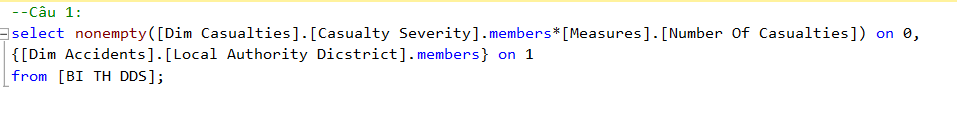
Thuộc tính chiều: Casual\_Severity (dim\_Casualties), Local\_Authority\_District (dim\_Accidents)

Kết quả thống kê tổng quát:

Table

Description automatically generated

MDX:



Kết quả:

Table

Description automatically generated

Câu 2: Thống kê số lượng nạn nhân theo Mức Độ Nghiêm Trọng ở các Địa Phương (Local\_Authority\_(District)) theo các Quý trong từng năm.

OLAP:

Measure: Number or casualties

Thuộc tính chiều: Casual\_Severity (dim\_Casualties), Local\_Authority\_District (dim\_Accidents), Quarter (dim\_Date), Year (dim\_Date)

Kết quả thống kê tổng quát:

Table

Description automatically generated

Kết quả thống kê theo năm 2011

Table

Description automatically generated

MDX:

Graphical user interface, application

Description automatically generated

Kết quả:

Graphical user interface, table

Description automatically generated

Câu 3: Thống kê số lượng người tử vong theo Giới Tính, Loại Nạn Nhân (Casualty

Type) và Nhóm Tuổi (Age\_Band\_of\_Casualty) theo các năm.

OLAP:

Measure: number of Casualties

Thuộc tính chiều: Casual\_Severity (Fatal) (dim\_Casualties), Casual\_type (dim\_Casualties), Sex of Casualty (dim\_Casualties), Age\_Band\_of\_Casualty (dim\_Casualties), Year (dim\_Date)

Kết quả thống kê tổng quát:

Table

Description automatically generated

Kết quả theo Male:

Table

Description automatically generated

Kết quả theo Female:

Graphical user interface, table

Description automatically generated

MDX Query:

A screenshot of a computer

Description automatically generated with medium confidence

Kết quả:

Female:

Table

Description automatically generated

Male:

Table

Description automatically generated

Câu 4: Thống kê số lượng TNGT theo Mức Độ Nghiêm Trọng và Thời Điểm Trong Ngày (Morning: 5am-12pm, Afternoon: 12pm-5pm, Evening: 5pm-9pm, Night: 9pm-5am) trong các năm.

OLAP:

Measure: Fact Accident count

Thuộc tính chiều: Accident\_Severity(dim\_Accidents), Year (dim\_Date), Time\_Band (dim\_Accidents)

Kết quả thống kê tổng quát:

Table

Description automatically generated

MDX:

Graphical user interface, application, Word

Description automatically generated

Kết quả:

Graphical user interface, table

Description automatically generated

Câu 5: Thống kê số lượng TNGT theo Mức Độ Nghiêm Trọng, Vùng

(Urban\_or\_Rural\_Area), và Kiểu Đường (Road Type) trong các năm.

Measure: Fact Accident count

Thuộc tính chiều: Accident\_Severity(dim\_Accidents), Year (dim\_Date), Road Type (dim\_Accidents), Urban\_or\_Rural\_Area (dim\_Accidents)

Kết quả thống kê tổng quát:

Table

Description automatically generated

MDX:

Text

Description automatically generated

Kết quả:

Table

Description automatically generated

Câu 6: Thống kê số lượng nạn nhân theo Mức Độ Nghiêm Trọng, Loại Nạn Nhân

(Casualty Type) và Độ Tuổi trong các năm

OLAP:

Measure: number of Casualties

Thuộc tính chiều: Casualties\_Severity (dim\_Casualties), Casualties\_Type (dim\_Casualties), Age\_Band (dim\_Casualties), Year (dim\_Date)

Kết quả thống kê tổng quát:

Graphical user interface, application, table

Description automatically generated

MDX:

Graphical user interface, text, application

Description automatically generated

Kết quả:

Table

Description automatically generated

Câu 7: Tổng hợp số lượng tai nạn theo Mục Đích Hành Trình (Journey Purpose) và

Loại Phương Tiện (Vehicle\_Type)

Measure: Accident\_Count, Number of Vehicle

Thuộc tính chiều: Journey Purpose (dim\_Vehicle), Vehicle\_Type (dim\_Vehicle)

Graphical user interface, table

Description automatically generated

MDX:

Graphical user interface, application, Word

Description automatically generated

Kết quả:

Table

Description automatically generated

Câu 8: Tạo thêm thuộc tính build-up Road

Text

Description automatically generated

Câu 9: Thống kê số lượng tai nạn theo Mức Độ Nghiêm Trọng, Loại Phương Tiện

(Vehicle Type), Built-up Road trong các năm.

Measure: number of vehicle, Accident\_Count

Thuộc tính chiều: Journey Purpose (dim\_Vehicle), Vehicle\_Type (dim\_Vehicle), Build-up Road (dim\_accident), Year (dim\_Date)

Kết quả thống kê tổng quát:

Graphical user interface, application, table

Description automatically generated

MDX:

Text

Description automatically generated

Kết quả:

Table

Description automatically generated

Mining:

Case Study: Dựa trên data có sẵn, dự đoán xem nếu có tai nạn thì mức độ nghiêm trọng của nó sẽ là như thế nào

1. Chuẩn bị dữ liệu:

Tạo một bảng chứa thông tin nạn nhân, nơi ở, giới tính, tuổi đặt tên là Prospective\_Casualties để làm input table

Bảng Prospective\_Casualties:

Table

Description automatically generated

Tạo view chứa thông tin các tại nạn và đặt tên là Targer\_Accident

Graphical user interface, table

Description automatically generated

Tạo một data source view mới và add 2 nguồn này vào data source

2. Chọn thuật toán sử dụng:

- Ta sẽ sử dụng các dữ liệu của nguồn để dự đoán mức độ nghiêm trọng của tai nạn nên ta sẽ sử dụng thuật toán Decision Tree

- Thuật toán Decision tree sẽ áp dụng tốt hơn trong Case Study này là vì:

+ Vì một vụ tai nạn sẽ phụ thuộc nhiều yếu tố nhiều lớp nên việc sử dụng thuật toán sẽ cho ra các quy tắc để dự đoán tốt hơn

3. Build Mô hình và kết quả dự đoán:

a. Mô hình cây Decision:

Diagram

Description automatically generated

b. Kết quả mining và predict:

Table

Description automatically generated

# Link Github

<https://github.com/wgslancer/TTKD-03>

# Link Drive nhóm

<https://drive.google.com/drive/folders/1EL_8DOQjGkZD7g1esKOoGhQ7Cq91rfkD?usp=sharing>