KHOA CÔNG NGHỆ THÔNG TIN
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
227 Nguyễn Văn Cừ, Phường 4, Quận 5, TP.HCM
Điện thoại: (08) 38.354.266 – Fax: (08) 38.350.096

CSC12107 – HTTT PHỤC VỤ TRÍ TUỆ KINH DOANH

ĐỒ ÁN MÔN HỌC

DOANMH: XÂY DỰNG VÀ KHAI THÁC KDL

I. Thông tin chung

Mã số bài tập:	DOANMH
Thời lượng dự kiến:	10 tuần
Deadline nộp bài:	31/12/2021
Hình thức:	Bài tập nhóm
Hình thức nộp bài:	Nộp qua Moodle môn học
GV phụ trách:	Nguyễn Thị Như Anh, Tiết Gia Hồng, Hồ Thị Hoàng Vy
Thông tin liên lạc với GV:	ntnanh@fit.hcmus.edu.vn tghong@fit.hcmus.edu.vn hthvy@fit.hcmus.edu.vn

II. Chuẩn đầu ra cần đạt

Bài tập này nhằm mục tiêu đạt được các chuẩn đầu ra sau:

- G3.3 Thiết kế lược đồ chuẩn hoá, đa chiều (sao, bông tuyết) dựa vào dữ liệu hệ thống tác vụ và yêu cầu phân tích từ tình huống cho trước
- G5.1 Triển khai quy trình ETL để rút trích dữ liệu từ nhiều nguồn, biến đổi, làm sạch dữ liệu, nạp dữ liệu vào KDL sử dụng công cụ SSIS
- G5.2 Xây dựng KDL đa chiều sử dụng công cụ SSAS và giải thích được lựa chọn phép toán OLAP phù hợp đối với 1 số yêu cầu phân tích.
- G5.3 Sử dụng một số công cụ biểu diễn dữ liệu (SSRS, powerBI, excel…) để biểu diễn kết quả phân tích, khai thác được (report, dashboard…)
- G5.4 Sử dụng SSAS và áp dụng các kỹ thuật mining tích hợp để thực hiện khai thác dữ liệu từ KDL xây dựng được.

III. Các yêu cầu & quy định chi tiết cho bài nộp

Xây dựng và phân tích dữ liệu về các vụ tai nạn giao thông ở UK trong khoảng 3-4 năm.

- **Mô tả dữ liệu:** Mô tả ý nghĩa các thuộc tính của các nguồn dữ liệu sau (chỉ cần mô tả các thuộc tính cần thiết cho đồ án):
  - Dữ liệu UK Car Accidents 2005 - 2015 (SV chỉ lấy dữ liệu 3-4 năm, hoặc lấy dữ liệu từ 2011-2014 được cung cấp sẵn): *https://www.kaggle.com/silicon99/dft-accident-data/discussion/28970?fbclid=IwAR1BvAiy8mEMy01XXAKtxLkX7Kx3kwPt3c3EYhwoxlWq5psikSAB2mVlF8A*

- Dữ liệu LSOA-Postcode mapping: *https://geoportal.statistics.gov.uk/datasets/postcode-to-output-area-to-lower-layer-super-output-area-to-middle-layer-super-output-area-to-local-authority-district-august-2021-lookup-in-the-uk/about*

- Dữ liệu UK-Postcodes: *https://github.com/academe/UK-Postcodes/blob/master/postcodes.csv*

- **Thiết kế kho dữ liệu (KDL), tổng hợp, nạp dữ liệu các nguồn vào KDL và thiết kế và xây dựng Cube**: Gợi ý:

  - Đối với England và Wales: mapping các nguồn dữ liệu trên để lấy giá trị xây dựng Geography dimension với phân cấp chiều như sau: Country > Region > County > Town City
  - Đối với Scotland và North Ireland, SV tự đề xuất cách để tạo giá trị cho Geography dimension
  - Chuyển đổi dữ liệu ngày tháng sao cho có thể tạo được Date dimension với phân cấp chiều: Year > Quarter > Month > Day
  - Xác định và thiết kế các phân cấp chiều khác để đáp ứng yêu cầu OLAP và Report

- **OLAP và Report**:

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. Bảng tham khảo:

Number

Local authority area	LA code	Killed	Seriously injured	Killed or seriously injured	Slightly injured	All casualties
Hertfordshire	E10000015	24	380	404	3,068	3,472
Hillingdon	E09000017	6	60	66	903	969
Hounslow	E09000018	9	58	67	939	1,006
Isle of Wight	E06000046	5	78	83	339	422
Isles of Scilly	E06000053	0	0	0	1	1
Islington	E09000019	2	87	89	885	974
Kensington and Chelsea	E09000020	4	48	52	656	708
Kent	E10000016	54	578	632	5,167	5,799
Kingston upon Hull, City of	E06000010	1	104	105	891	996
Kingston upon Thames	E09000021	3	26	29	353	382

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.**

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. Bảng tham khảo:

Number of casualties

Female		2010-14 average[1]	2008	2009	2010	2011	2012	2013	2014	2015
Pedestrians	0 to 4[2]	75	86	76	66	92	76	68	72	61
	5 to 7	85	83	80	82	112	77	75	77	78
	8 to 11	167	168	163	196	188	162	145	146	148
	12 to 15	234	305	297	269	250	237	210	205	208
	16 to 19	154	217	182	153	186	170	143	116	138
	20 to 24	153	180	159	161	158	156	143	149	143
	25 to 59	679	745	651	599	663	736	678	718	676
	60 to 64	106	111	117	96	109	108	101	114	92
	65 to 69	103	94	96	82	92	106	115	118	128
	70 to 74	115	133	115	105	122	114	104	131	107
	75 to 79	130	145	120	124	120	149	120	137	129
	80 and over	255	326	287	257	263	232	246	275	250
	All age groups[3]	2,280	2,649	2,376	2,215	2,388	2,344	2,178	2,276	2,178
Pedal cyclists	0 to 4[2]	2	1	1	2	2	2	0	2	0
	5 to 7	7	0	11	10	9	7	7	2	6
	8 to 11	21	28	18	30	27	21	14	15	12
	12 to 15	21	20	25	25	23	20	20	18	19
	16 to 19	25	22	15	21	26	23	26	27	22
	20 to 24	52	51	56	36	60	46	53	64	48
	25 to 59	384	276	295	321	364	410	402	424	397
	60 and over	56	52	46	69	52	49	44	64	63
	All age groups[3]	575	459	471	524	571	581	576	621	575
Motorcycle riders	Under 16	0	2	1	0	0	0	1	0	0
50cc and under	16	12	15	11	14	15	11	12	10	9
	17	4	8	6	9	1	3	3	4	4
	18	4	7	2	3	4	3	4	4	2
	19	2	3	5	2	2	2	1	2	2
	20 to 24	9	9	4	6	13	8	10	6	8
	25 to 59	25	36	24	19	24	39	20	23	17
	60 and over	5	6	7	5	6	9	4	1	3

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.

5. Thống kê số lượng TNGT theo **Mức Độ Nghiêm Trọng, Vùng** (Urban_or_Rural_Area)**, và Kiểu Đường** (Road Type) trong các năm.

6. Thống kê số lượng nạn nhân theo **Mức Độ Nghiêm Trọng, Loại Nạn Nhân** (Casualty Type) **và Độ Tuổi** trong các năm, **Độ Tuổi** được định nghĩa như sau:
   - Children: 0-15
   - Young adult: 0-17
   - Adult: 18-59
   - 60 and over: 60-...

Bảng tham khảo:

	Killed		Seriously injured		Slightly injur
	Number	% change	Number	% change	Number	c
Pedestrians					
Children: 0-15 years	25	-14	1,258	-7	5,034	
Young people: 0-17 years	32	-20	1,411	-6	5,796	
Adults: 18-59 years	203	-6	2,276	1	9,826	
60 and over	173	-9	1,181	-6	2,659	
All casualties[1]	408	-9	4,940	-2	18,713	
Pedal cyclists					
Children: 0-15 years	6	0	272	0	1,651	
Young people: 0-17 years	6	0	347	-9	2,178	
Adults: 18-59 years	69	-8	2,525	-5	12,175	
60 and over	25	-22	333	-1	806	
All casualties[1]	100	-12	3,239	-5	15,505	
Car occupants					
Children: 0-15 years	19	6	315	-1	6,681	
Young people: 0-17 years	42	27	555	-3	9,248	
Adults: 18-59 years	480	-6	5,492	-2	79,568	
60 and over	232	-9	1,755	-3	12,902	
All casualties[1]	754	-5	7,888	-2	103,065	
Motorcycle users	365	8	5,042	-5	14,511	

7. Tổng hợp số lượng tai nạn theo **Mục Đích Hành Trình** (Journey Purpose) **và Loại Phương Tiện** (Vehicle_Type)**.** Bảng tham khảo:

Number of vehicles/*percentage*

Journey purpose		Pedal cycle	Motorcycle	Car	Bus or coach	Vans / Light goods vehicles	Heavy goods vehicles	All vehicles[1]
Work	No. of vehicles	1,125	1,806	19,192	4,608	6,480	5,250	39,785
	Percentage	*6*	*9*	*10*	*86*	*47*	*81*	*15*
Commuting	No. of vehicles	3,115	3,366	19,054	23	1,260	85	26,966
	Percentage	*16*	*16*	*10*	*0*	*9*	*1*	*10*
Taking Pupil to School	No. of vehicles	54	23	2,484	42	26	1	2,634
	Percentage	*0*	*0*	*1*	*1*	*0*	*0*	*1*
Pupil Riding to School	No. of vehicles	457	110	239	3	3	2	817
	Percentage	*2*	*1*	*0*	*0*	*0*	*0*	*0*
Other / Unknown	No. of vehicles	14,686	15,690	147,888	705	6,104	1,131	187,619
	Percentage	*76*	*75*	*78*	*13*	*44*	*17*	*73*
Total	No. of vehicles	19,440	20,996	188,872	5,381	13,876	6,470	257,845
	Percentage	*100*	*100*	*100*	*100*	*100*	*100*	*100*

8. Tạo thêm thuộc tính **Built-up Road** trong table Accidients. **Built-up Road** có 2 giá trị:
   - Built-up road: Nếu tốc độ giới hạn (**Speed Limit**) dưới 50 mph
   - Non Built-up road: Nếu tốc độ giới hạn từ 50 mph

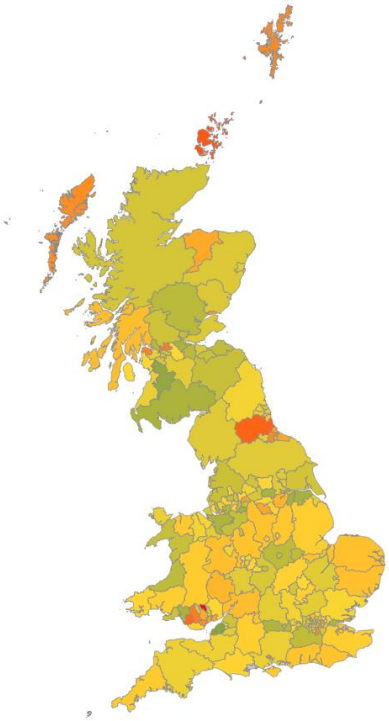
9. Thống kê số lượng tai nạn theo **Mức Độ Nghiêm Trọng, Loại Phương Tiện** (Vehicle Type)**, Built-up Road** trong các năm.

10. Sinh viên tự thiết kế những bảng thống kê khác để có thêm nhiều chiều đánh giá TNGT ở UK.

11. Định nghĩa fact **Variance** để tính mức độ tăng giảm của TNGT theo đơn vị phần trăm qua các năm.

12. Xây dựng đồ thị/ biểu đồ cho các bảng thống kê ở trên.

13. Dùng regional map để biểu diễn trực quan (bằng màu sắc) số lượng TNGT ở các vùng trong năm. Ví dụ tham khảo:



- **Data Mining:** Gợi ý:
  - Sử dụng mô hình dự đoán mức độ nghiêm trọng của các tai nạn
  - Sinh viên có thể đề xuất ứng dụng một trường hợp bất kỳ, lý giải thuật toán sử dụng, vì sao, kết quả như thế nào,…

- **Kết luận chung:**

IV. Cách đánh giá

- Vấn đáp giữa kỳ: ETL process (data flow, data cleaning, ETL data from source to DW)
- Vấn đáp cuối kỳ: Project hoàn chỉnh (khai thác KDL với report, olap, mining, tạo job tự động định kỳ thực hiện ETL)

V. Tài liệu tham khảo

- https://data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data

VI. Các quy định khác

- Sinh viên làm việc nhóm và post mã nguồn lên Github
- Project bao gồm:
  - File báo cáo gồm:
    - Thông tin thành viên
    - Chi tiết phân công công việc, % hoàn thành
    - Xuất report từ github
  - Nội dung chính:
    - Phân tích và thiết kế KDL (NDS, DDS)
    - Phân tích quá trình ETL dữ liệu (làm sạch, chuyển đổi, tích hợp dữ liệu, …)
    - Khai thác KDL (OLAP, Report, Mining)
  - Source:
    - script tạo csdl NDS, DDS
    - Project ETL, mining…
- Đánh giá:
  - Giáo viên đánh giá điểm tổng cho nhóm, nhóm tự xác định tỉ lệ điểm của mỗi thành viên tùy theo mức độ đóng góp vào đồ án.
- Kế hoạch nộp đồ án dự kiến:
  - Giữa kì: khoảng tuần 7-8
  - Cuối kì: khoảng tuần 14-15