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 – INFORMATION SYSTEM FOR BUSINESS INTELLIGENCE

PROJECT

BUIDING AND MINING DATA WAREHOUSE

I. General information

Assignment ID:	PROJECT
Estimated duration:	10 weeks
Submission deadline:	31/12/2021
Assignement type:	Student Group
Submission chanel:	Moodle
Teachers:	Nguyễn Thị Như Anh, Tiết Gia Hồng, Hồ Thị Hoàng Vy
Contacts:	ntnanh@fit.hcmus.edu.vn tghong@fit.hcmus.edu.vn hthvy@fit.hcmus.edu.vn

II. Learning outcomes

This assignment is to gain the following outcomes:

- G3.3 Design a Star or Snowflake data model diagram through the Multidimensional Design from analytical business requirements and OLTP system
- G5.1 Deploy the ETL procedure to extracting data from disparate databases and data sources, and then transforming the data for effective integration into a data warehouse using SSIS tool
- G5.2 Operate the basic OLAP technologies using SSAS tool.
- G5.3 Create dashboard and other visualizations to analyze and communicate the data from DW using SSRS or excel...
- G5.4 Applying the data mining algorithms in Analysis Services to your data.

III. Requirements and submission rules

Build and analyze data about car accidents in the UK over 3-4 years.

- **Data Description:** Describe meaning of the properties of the following data sources (only describe the properties necessary for the project):
  - UK Car Accidents 2005 - 2015 data (Students only takes data over 3-4 years, or the provided 2011-2014 data): *https://www.kaggle.com/silicon99/dft-accident-data/discussion/28970?fbclid=IwAR1BvAiy8mEMy01XXAKtxLkX7Kx3kwPt3c3EYhwoxlWq5psikSAB2mVlF8A*

- LSOA-Postcode Mapping data: *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*

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

- **Design data warehouse (DW), synthesize, load data from the sources into DW, then design and build Cube: Suggestions:**

  - For England and Wales: map the above data sources to get the values for building Geography dimension with dimensional hierarchy as follows: Country > Region > County > Town City
  - For Scotland and North Ireland, students need to suggest solutions to create values for Geography dimension
  - Transform the datetime data to create the Date dimension with dimensional hierarchy: Year > Quarter > Month > Day
  - Define and design other dimensional hierarchies to meet OLAP and Report requirements

- **OLAP and Report**:
  1. Report the number of calsulities by **Severity** (Fatal, Serious, Slight) in the **Local Authority Districts** over years. Reference:

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. Report the number of calsulities by **Severity** (Fatal, Serious, Slight) in the **Local Authority Districts** by **Quarters in years.**

  3. Report the number of faltal calsulities by **Gender, Casualty Type and Age Band** over years. Reference:

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. Report the number of accidents by **Severity** and **Time of Day** (Morning: 5am-12pm, Afternoon: 12pm-5pm, Evening: 5pm-9pm, Night: 9pm-5am) over years.

5. Report the number of accidents by **Severity, Urban or Rural Area** and **Road Type** over years.

6. Report the number of calsulities by **Severity, Casualty Type** and **Age Group** over years, **Age Group** is defined as below:
   - Children: 0-15
   - Young adult: 0-17
   - Adult: 18-59
   - 60 and over: 60-...

Reference:

	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. Report the number of accidents by **Journey Purpose** and **Vehicle Type.**
   Reference:

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. Create a new attribute **Built-up Road** in Accidients table. **Built-up Road** may have 2 values:
   - Built-up road: if **Speed Limit** below 50 mph
   - Non Built-up road: if **Speed Limit** above 50 mph

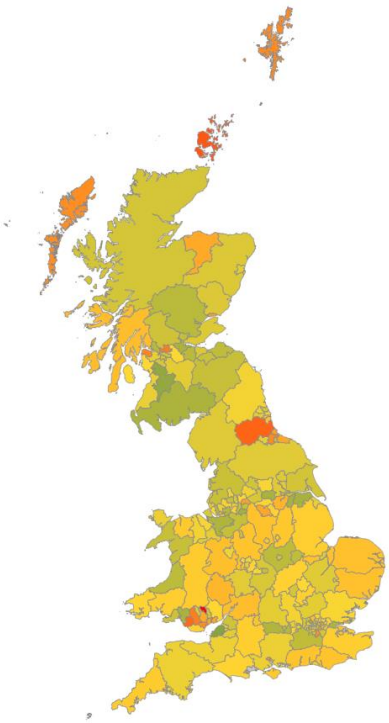
9. Report the number of accidents by **Severity, Vehicle Type, Built-up Road** over years.

10. Students design other reports about UK car accident.

11. Define fact **Variance** to calculate the increase and the decrease of the number of car accidents in percent over years.

12. Build graphs/charts for the above reports.

13. Use regional map to visually represent (by color) the number of car accidents in regions during a year. Reference:



- **Data Mining:** Suggestion:
  - Using models to predict the severity of accidents
  - Students propose applications of any case, explain the algorithm used, why, how the results are, etc.

- **Conclusion:**

IV. Assessment

- Midterm Q&A: ETL process (data flow, data cleaning, ETL data from source to DW)
- Final Q&A: Completed project (mining DW with reports, olap, mining, periodical automatic job creation to perform ETL)

V. References

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

VI. Other rules

- Students work in groups and post the source code on Github
- Project includes:

  - The report file includes:
    - Members Information
    - Details of work assignment, % tasks completed
    - Export report from github
  - Main content:
    - Analysis and design of databases (NDS, DDS)
    - Data ETL process analysis (cleaning, transformation, data integration, ...)
    - Data mining (OLAP, Report, Mining)
  - Source:
    - Script to create database NDS, DDS
    - Project ETL, mining…
- Assessment:
  - The teacher evaluates the total score for each group, the group determines the percentage of each member's score depending on the level of contribution to the project.
- Project submission plan:
  - Midterm: around week 5-6
  - Final: around week 11-12