Coursework Description

Overview

This individual coursework counts 15% of your final mark. The purpose of this coursework is to normalise real datasets, identify interesting research questions and apply dimensional modelling to answer interesting research questions.

Coronavirus has had a huge impact in our lives. As a result there are lots of datasets available containing different kinds of data. In this coursework you will work with some of these datasets to answer interesting research questions. In particular, you are asked to work with 2 datasets: a coronavirus dataset from the EU Open Data portal and a dataset with public health and social measures (PHSM) implemented by countries collected by the WHO.

Datasets

To facilitate your work the datasets have been already imported into a database and a backup file is available on KEATS. When you restore the backup you will see 2 tables (corona and PHSM) corresponding to the 2 datasets.

Together with the datasets you will also find some documents providing information about the PHSM dataset.

What you need to submit

On KEATS you will find a folder containing the template files that you need to complete for your assignment. You need to submit your edited coursework files as a zip file.

We will evaluate these SQL files using MySQL server. *Any file that is missing, renamed, or does not run will result in 0 points for that section.*

Coronavirus data from the EU Open Data Portal (30%)

The EU Open Data Portal provides access to an expanding range of data from the European Union (EU) institutions and other EU bodies.
The Coronavirus dataset contains the latest available public data on COVID-19 including a daily situation update formed by the number of COVID-19 cases and deaths, based on reports from health authorities worldwide.

1. Normalization (worth 10%). In the *corna2BD.sql* file, include the SQL commands for normalising the coronavirus dataset into a relational database.
2. Research questions (worth 5%). In the *coronaResearch.sql file,* write as comments 2 different research questions in plain text. The importance and relevance of the questions will be taken into account to assess this task. An

example of a simple and not very interesting question is the calculation of the death counts per country and continent.

3. Data Mart (worth 10%). In the *corona2DM.sql file,* include the SQL commands for creating a data mart (star or snowflake schema) to answer your research questions questions.

4. Queries to answer research questions (worth 5%). In the *coronaQueries.sql* file, include the SQL queries that answer your research questions.

PHSM dataset from the WHO (30%)

Public health and social measures (PHSMs) are measures or actions by individuals, institutions, communities, local and national governments and international bodies to slow or stop the spread of an infectious disease, such as COVID-19. Since the start of the COVID-19 pandemic, a number of organizations have begun tracking implementation of PHSMs around the world, using different data collection methods, database designs and classification schemes. The WHO's PHSM has combined these datasets together, using a common taxonomy and structure, into a single, open-content dataset for public use.

1. Normalization (worth 10%). In the *PHSM2BD.sql* file, include the SQL commands for normalising the coronavirus dataset into a relational database.

2. Research questions (worth 5%). In the *PHSMResearch.sql file,* write as comments 2 different research questions in plain text. The importance and relevance of the questions will be taken into account to assess this task. An example of a simple and not very interesting is the number of different measures taken by each country.

3. Data Mart (worth 10%). In the *PHSM2DM.sql file,* include the SQL commands for creating a data mart (star or snowflake schema) to answer your research questions questions.

4. Queries to answer research questions (worth 5%). In the *PHSMQueries.sql* file, include the SQL queries that answer your research questions.

Integration of Coronavirus data and PHSM (40%)

1. Research questions needing both Data Marts (worth 15%). In the *integrationResearch.sql* file, you need write 2 research questions (as comments in plain text) that combine data from the 2 data marts that have been previously built as comments. The importance and relevance of the questions will be taken into account to assess this task. For example, a research question may be to determine the number of cases the day in which the measures where applied in a particular country.

2. Integration of both DM (worth 15%). In the *integrationDM.sql* file, you should include the SQL commands for integrating the 2 data marts. Note that data mart integration is should be done by means of common dimension tables. Do not try to link the fact tables directly.
3. Queries to answer research questions (worth 10%). In the *integrationQueries.sql* file, you need to include the SQL that answer the research questions stated in the point above.

FAQ

If I cannot think of any research question for some of the datasets, can I use the example research questions in this description?

Yes, but that will mean that you will not receive any marks for these questions.

Do I need to include the commands for normalising and creating the individual data marts in the integrationDM.sql file?

No, I will assess your integration section just after assessing the coronavirus and PHSM parts.

‹ 6.18 General discussion - week 6

Datasets ›