Assessment 3 and Assessment 4

|  |  |
| --- | --- |
| ***Deadline:*** | Hand in by midnight April 27 2024 |
| ***Work*** | If you are doing this project with someone else, state this clearly in your submission, together with a document that clearly specifies which tasks each person completed in  the project. |
| ***Purpose:*** | Learn how to implement the full process of data acquisition, data wrangling, data integration, data persistence using SQLite, and data analysis using Python. |

## Assessment 3 and 4 overarching outline:

The goal of these projects is the implementation of a full data analysis workflow using python with the combination of SQLite database persistence.

You are asked to preferably choose a problem domain that is aligned with your specialization within select a domain of interest to you. You may re-use some of the datasets from the previous assignment. Research what kinds of data sources are available for your selected domain. Subsequently, you are asked to (1) formulate questions that you would like answered, (2) acquire datasets from at least two different sources (at least one source must be dynamic, i.e. is web-scraped or is retrieved from a web API), (3) wrangle the data into an usable format and perform EDA, (4) integrate datasets into one, (5) persist the data into a SQLite relational database with a suitable schema, (6) perform group-by queries, pivot tables, cross-tabulation of the data to answer your research questions, together with a rich set of visualizations.

Links to various dataset and web API repositories are provided on Stream. The analysis workflow you are asked to perform is illustrated in the diagram.

below:

A white background with black text

Description automatically generated

## Assessment 4 Requirements:

Your research report must be in a Jupyter Notebook format and thus executable and repeatable. Clearly introduce your problem domain, articulate your research questions and provide an executive summary at the beginning. Follow the provided Jupyter notebook template.

You must document and explain the reasoning behind the coding steps you are taking and provide explanations of all your graphs and tables as is appropriate. Make sure you label all aspects of your graphs.

The activities listed under the five stages in the workflow diagram above are a guide only. This means that operations like group-by statements as well as pivot tables could be a part of the ‘Data Wrangling’ phase as EDA, and not only a part of the data analysis phase.

## Assessment 4 Marking criteria:

Marks will be awarded for different components of the project using the following rubric:

|  |  |  |
| --- | --- | --- |
| **Component** | **Marks** | **Requirements and expectations** |
| Data Acquisition | 20 | * Diversity of sources (at least one must be dynamic – full marks for using both APIs and web scraping – penalties will be applied for re-using examples from class) * Appropriate use of merging and concatenation. |
| Data Wrangling and EDA | 30 | * Quality of your EDA * Appropriate use of visualisations * Thoroughness in data cleaning * Use of user-defined functions |
| Data Analysis | 35 | * Quality of the research questions being asked * Diversity of techniques used to answer and present them * Clear and structured presentation of findings * Interpretation and communication of findings and visualisations |
| Originality and challenge | 15 | * Originality in problem definition and approach to the analysis * Creativity in problem solving * The degree of challenge undertaken |
| **BONUS** |  |  |
| Big Data Processing Techniques | 5 | * Demonstration of out-of-core processing * Analysis of query performance issues and optimisations where necessary |

## Assessment 3 Specific Requirements:

Once you have completed the above components, your task now is to design a database (DB) schema that represents all the data that you have acquired from multiple sources in a normalised form, and to populate it using SQLite, thus achieving full data persistence.

The project requirements are as follows:

* create a separate Jupyter Notebook for these tasks
* create a simple DB schema document that shows the tables (aim for around half a dozen), their attributes and relationships that depict your design; use free software like DBSchema to create a diagram and embed this image into the notebook
* create an image file from the schema DB design document and embed it into your notebook
* describe your DB schema at a high level
* write all the database schema code for creating the necessary tables for SQLite DBs
* read in all the data that you have prepared in the above project and which you have stored in various file formats (.csv and/or .xlsx) and populate your tables from the notebook
* perform some analysis that requires extracting data from your DB; write at least six queries that require various table joins on your DB; these queries can replicate or be based on some of the analysis that you performed in the above project. You may also include some visualisations in the notebook.
* create at least two DB Views which encapsulated queries from above and test them

## Assessment 3 Marking criteria:

Marks will be awarded for different components of the project using the following rubric:

|  |  |  |
| --- | --- | --- |
| **Component** | **Marks** | **Requirements and expectations** |
| Schema Definition | 35 | * design of a DB schema document and its explanation * creating half a dozen normalised tables that capture all the data * use of correct data types for attributes * definition of primary and foreign keys where appropriate * definition of indexes where appropriate * definition, implementation, and explanation of constraints where necessary |
| DB Population | 20 | * automation of reading files from flat-files and writing data into SQLite tables * performing checking that the data has been persisted in the SQLite DB |
| SQLite Queries | 35 | * complexity of queries (these should be much more than simple SELECT statements) * diversity of queries * readability and structure of the SQL code * explanation of the queries and results |

|  |  |  |
| --- | --- | --- |
| DB Views | 10 | * creation of two DB Views * testing out the views |

**Hand-in**: Submit your zipped notebook(s) file together with your final datasets and SQLite database.