Data Analysis and Integration

Project

In this project, we will be using two datasets by E-REDES available in the Open Data Portal:

- Dataset A: **Number of active energy contracts by meter type**
- Dataset B: **Monthly consumption by municipality**

Both datasets are available for download as a CSV file on the following website:

- https://e-redes.opendatasoft.com/explore/

These datasets contain real-world data about the energy distribution network in Portugal.

Tasks

1. You will notice that dataset A provides the number of contracts (CPEs) by time, location and meter type. On the other hand, dataset B provides the energy consumption (kWh) by time, location and voltage level. The first task is to build a transformation that provides the percentage of smart meters and the average energy consumption per contract. These need to be calculated by municipality in June 2024. The output should be a CSV file with three columns: Municipality, Percentage of smart meters, Consumption per contract.

2. Based on the output from the previous task, use DataCleaner to generate a scatter plot to study the correlation between the percentage of smart meters (%, x-axis) and energy consumption per contract (kWh/CPE, y-axis). From the plot, identify whether there is a trend between these two variables.

3. Create a transformation to identify the correct mapping of regions ('districts' in Portuguese) in dataset A to regions in dataset B. The transformation should be based on a duplicate detection approach using a *string matching* measure and a threshold. The output should be a CSV file with two columns (District_A and District_B) and eighteen rows (not including the header). There should be no false positives or false negatives in the results. Regions ('districts') should be sorted alphabetically.

4. Based on dataset B, create an SQL schema creation script to create the tables for a data warehouse with two dimensions: time (with year, season, month) and location (with region, municipality, parish). The season will be summer, autumn, winter or spring. The fact table should have energy consumption and percentage of smart meters as the measures. Use this script to create the data warehouse schema in MySQL.

5. Develop the transformation or transformations to populate the dimension tables.

6. Develop the transformation to populate the fact table.

7. Use Pentaho Schema Workbench (PSW) to define an OLAP cube based on this data warehouse. The result should be saved as an XML file.

8. Use Pentaho Server and Saiku Analytics to perform an analysis on the data cube. In particular, analyze the energy consumption by region (i.e., 'district') and year, but only for years with complete data from January to December. **For each region, check the following**:

a. Has the consumption and the percentage of smart meters increased or decreased from one year to the next?
b. What is the influence of the season on consumption?
c. Is the percentage of smart meters having an impact on electricity consumption?

Deliverables

Prepare a PowerPoint (or similar) slide presentation with the following elements from each task:

1. Present a screenshot of the entire transformation. Present screenshots of the configuration window and of the preview window for each step. Present a screenshot of the output file.

2. Present a screenshot of the analysis job (i.e. the sequence of steps) that you developed with DataCleaner. Present a screenshot of the configuration window for each step. Present a screenshot of the plot that you got in the analysis results. Is there a correlation between these variables? If yes, how strong is it?

3. Present a screenshot of the entire transformation. Present screenshots of the configuration window and of the preview window for each step. Present a screenshot of the output file.

4. Present a screenshot with the entire contents of the SQL script.

5. For each transformation that you develop, present a screenshot of the entire transformation, screenshots of the configuration window and of the preview window for each step, and a screenshot of the output table (only some rows, if it is not possible to show them all).

6. Present a screenshot of the entire transformation, screenshots of the configuration window and of the preview window for each step, and a screenshot of the output table (only some rows, if it is not possible to show them all).

7. Present a screenshot of PSW with the OLAP cube definition fully expanded on the left pane, and a screenshot with the entire contents of the corresponding XML file.

8. Present screenshots of all the analysis queries and results that you generate with Saiku Analytics. Indicate, in particular:

    a. The desired increase/decrease for each district, and the analysis of the impact of smart meters;
    b. The influence of the seasonality on consumption;
    c. The impact of smart meters on consumption.

Optionally, you may include additional text or comments in the slides to help clarify, or to draw conclusions from, the results that you are presenting.

Submission

The project must be delivered in a **Zip** a file named **submission-aid-GG.zip** (where **GG** is the group number), via Fénix system until 23h59 of the day of the submission deadline.

The **Zip** file must have the following structure:

Presentation-GG.pdf (where GG is the group number)	The presentation **pdf** where **GG** is the group number. **Check that the image quality is good enough for all screenshots to be readable.**
etl/	Folder containing all Pentahoo transformations and jobs
dw/	Folder containing all the SQL scripts to create the Data Warehouse
olap/	Folder containing OLAP cube definitions
output/	Folder where output files are stored

NOTE: Penalties will apply to groups that do not follow the submission instructions. Evaluation elements that are not found in the prescribed as above **will not be taken into account for grading purposes**. DO NOT wait until the last moment to submit. Late submissions will NOT be accepted.