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 has 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 goal of this task is to build a transformation that provides the number of contracts and the energy consumption per municipality, as of June 2023. The output should be a CSV file with three columns: Municipality, CPEs, kWh.

2. Based on the output from the previous task, use DataCleaner to generate a scatter plot in order to study the correlation between the number of contracts (CPEs, on the x-axis) and the energy consumption (kWh, on the y-axis). From the plot, identify the four topmost municipalities in terms of contracts and/or consumption.

3. Build a transformation to identify the correct mapping of districts in dataset A to districts 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 counting the header). There should be no false positives or false negatives in the results. The districts should be sorted alphabetically.

4. Based on dataset B, prepare an SQL script to create the tables for a data warehouse schema with three dimensions: time (with year, month), location (with district, municipality, parish) and voltage (with level). The fact table should have the energy consumption as a measure. 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 district and year, but only for years with complete data from January to December. For each district, check whether consumption has increased or decreased from one year to the next. Finally, delve deeper into the analysis and try to find out whether such increase/decrease is due to an increase/decrease at a low or at a high voltage level.

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. Identify the desired municipalities and their location in that plot.

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 the desired increase/decrease for each district, and the analysis according to voltage level, as requested.

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

Save your slide presentation as a PDF file.

Check that the image quality is good enough for all screenshots to be readable.

Submit the PDF file in Fénix before the project deadline.