Project report

Project name: C19 NRG consumption. Impact of Covid-19 on energy consumption in Baltic countries.
Team members: Anna Rutmane, Normunds Bērziņš
GitHub repository: https://github.com/nb3rz1ns/NRG_Project

Business understanding.

Identifying your business goals.

Background.

The outbreak of SARS-CoV-2 lead to a global pandemic COVID-19. Soon after the start of pandemic, situation in Baltic States became critical. February 2022 marked the peak of cases, with over 10 000 new cases reported daily. The governments of all three Baltic states imposed strict measures to curb the spread of the virus, including nation-wide emergency states, lockdowns, closing of educational institutions and places of entertainment, curfew during winter holidays, etc. The measures to control the spread of infection adapted by Baltic States vary in severity and timeline. Such measures during emergency states likely influenced energy consumption in all Baltic States and changed energy consumption patterns.

Business goals.

See how the global COVID-19 pandemic and associated countermeasures such as lockdowns influenced energy consumption in the Baltic states – Estonia, Lithuania, and Latvia. This may allow future energy consumption predictions in case of continuation of COVID-19 pandemic or a new pandemic or new countermeasures.

Business success criteria.

Recognizable trends and patterns in daily energy consumption data during the COVID-19 period. Accurate calculation of the percentage change in average daily energy consumption for each country between pre-COVID and COVID-19 periods. Establishing correlations between energy consumption trends and key COVID-19 events or restrictions in each country.

Assessing your situation.

Inventory of resources.

- Up-to-date data on COVID-19 parameters, including the number of reported cases in Estonia, Lithuania and Latvia.
- Dataset about energy consumption in Baltic states for relevant time period.
- Information about lockdown measures including closure and reopening dates of educational institutions, shops, entertainment places and other services.
- Data analysis and visualization tools using Python script.

Requirements, assumptions, and constraints.

The project must be completed by 11th of December, 2023, submission date for poster. There are no security obligations as the used data is freely available and the end the results of the project are not bound by any privacy laws. Legal obligations include presenting this project

on 14th of December, 2023 as to pass the course Introduction to Data Science (LTAT.02.002). Finished project and presentation must fit the criteria set by the course lecturers.

Risks and contingencies.

Risk: Incomplete or inaccurate data may compromise the accuracy of the analysis.
Contingency: Regularly validate and clean the data and establish protocols for handling missing or inconsistent data.

Risk: The complexity of modeling energy consumption patterns may lead to difficulties in interpretation or model errors.
Contingency: Conduct thorough model validation. Develop simple and interpretable models where possible.

Risk: Expansion of the project scope beyond its initial objectives may result in delays and resource overruns.
Contingency: Define and document the project scope. Regularly review project goals and objectives to ensure alignment. Try to extract the maximum of information from datasets that we already have before searching for additional tools.

Terminology.

Time series analysis – a specific way of analyzing a sequence of data points collected over an interval of time.
Regression analysis – a statistical method that allows to examine the relationship between two or more variables of interest.
Addition of relevant terms may be necessary over the course of project.

Costs and benefits.

We do not have any expenses related to the project, since we are using free datasets and free computing resources, and we are not consulting any experts in the field.

Benefits include acquiring insights about relationship of energy consumption and COVID-19 parameters, insights into the impact of school closures on energy consumption and contribution to public awareness regarding the interconnectedness of external events and energy consumption. Information acquired upon completing the project would benefit energy providers, policy makers and general public.

Defining your data-mining goals.

Data-mining goals.

Objective**:** Understand the temporal patterns and trends in weekly energy consumption during the COVID-19 pandemic.
Goals for Time Series Analysis:

- Identify and quantify the impact of COVID-19 on weekly energy consumption.
- Uncover any seasonality or recurring patterns in energy usage over time.

Goals for Regression Analysis:

- Determine the relationships between external factors (e.g., new covid cases, government restrictions) and weekly energy consumption.

- Develop regression models to predict energy consumption based on relevant variables.

Data-mining success criteria.

Time Series Analysis Success Criteria:

- Metric: Mean Absolute Percentage Error (MAPE) for Time Series models.
- Criterion: Achieve a MAPE below a predetermined threshold, indicating accurate prediction of weekly energy consumption trends during the pandemic.

Regression Analysis Success Criteria:

- Metric: Coefficient of determination (R-squared) for regression models.
- Criterion: Attain a high R-squared value, signifying the ability to explain a significant portion of the variance in weekly energy consumption through the identified external factors.

Data understanding

Gathering data.

Outline data requirements.

Reported COVID-19 cases over the pandemic.
Reported energy consumption in the Baltic State countries: Estonia, Lithuania, and Latvia.
COVID-19 related countermeasure event times and durations

Verify data availability.

COVID-19 cases were reported by more than 94 countries over the pandemic (some daily, some weekly). These datasets and compilations of them are freely available (we use a version of this from Kaggle).
Energy consumption in each of the three countries are freely available on various websites for each country.
COVID-19 related countermeasure event times and durations are not available in the form of a dataset, however these events are reported in news media and wiki sites, therefore are also available.

Define selection criteria.

Reported COVID-19 cases taken from a Kaggle dataset
https://www.kaggle.com/datasets/sandhyakrishnan02/latest-covid-19-dataset-worldwide/data
New cases, Countries (Latvia, Lithuania and Estonia) and Dates (from January 2020 to April of 2023) fields are to be used, however we took additional fields (such as ISO code, Continent, New deaths, etc.) when downloading the dataset to insure that during the project we are not missing any data that may be necessary or that may be mislabeled)

Energy consumption for each country taken from following websites:
For Estonian data: https://dashboard.elering.ee/en/system/with-plan/production-consumption?interval=minute&period=days&start=2023-11-08T22:00:00.000Z&end=2023-11-09T21:59:59.999Z
For Lithuanian data: https://www.litgrid.eu/
For Latvian data: https://www.ast.lv/lv/content/situacija-energosistema
Hourly energy consumption for each country as well as date-time (Date and time of measurements from 2018 to end of 2022) were taken for all 3 countries.

For COVID-19 related countermeasure events we scoured cited Wikipedia pages and news media pages. Types of events and dates for each country were noted in a PDF file.

Describing data.

COVID-19 dataset was uploaded to Kaggle by user SANDHYAKRISHNAN02 and was available to use in .csv format (links for this and other datasets available in section above). The dataset itself contains 67 columns of data; the specifics can be seen in the site. The data we plan to use: Date, Country and New Cases were available and extracted from this dataset. To reduce file size, only the data for 3 Countries – Latvia, Lithuania and Estonia – were extracted. The dataset is an ideal fit for this project as it contains the necessary data and the unnecessary parts of it can be omitted at acquiring stage.

Energy consumption data was provided by the author of project's idea, M. Sc. Neha Sharma, who acquired them from the sources listed before. We have three csv files for each Baltic

country. In each csv file there are two columns – Date Time and Consumption. Date Time column includes hourly energy consumption data starting from 1st January 2018 until 31st December 2022. Consumption column includes energy consumption in MW. Data is suitable for our project and is sufficient to reach the main goal of the project.

Data about COVID-19 lockdown is a descriptive data containing time periods and descriptions of specific measures and restrictions (when they were imposed and when they ended).

Exploring data.

Latvia and Estonia reported COVID cases daily, while Lithuania reported them weekly. To achieve a comparable results we had to recalculate number of daily COVID cases to number of weekly COVID cases for Latvia and Estonia.

Data Time column in energy consumption datasets were differently formatted – for Estonia and Latvia time data was presented in following format 1/1/2018 0:00, while for Lithuania it was in different format (01/01/2018 00:00:00). Everything was reformatted using pandas to_datetime function. Also, energy consumption hourly data was recalculated to weekly data in order to match COVID-19 cases dataset.

By the initial look, there is no correlation between the energy consumption and number of cases reported, but further analysis is required since the energy consumption in not only affected by COVID-19 cases, and other COVID-19 pandemic parameters should be taken into the account.

Verifying data quality.

The overall quality of data is very high. There are no missing datapoints, everything seems to be in order and most importantly the data are true reflection of the real-world construct (meaning the cases and energy consumption are true measurements, not calculated based on some parameters or estimates). The datasets we are looking at are not fully overlapping (energy consumption only up to 31st Dec 2022), however this is fine as there is no true relevancy to the available COVID data for 2023 (no major events and low cases). However, if we do deem energy consumption data for the non-overlapping period necessary, we can get it from the same sources as before. The data is also easy to use, accurate and reliable.

Planning your project

1. Obtaining and exploring data energy consumption and COVID cases datasets. Extract only necessary entries and perform data cleanup. During data clean-up reformatting and recalculations are performed. Merging of datasets over weekly dates. Estimated time: 7 hours. Performed by N. Berzins.
2. Obtaining COVID countermeasure event data by manually scouring the internet. Estimated time: 5 hours. Performed by A. Rutmane.
3. Set up GitHub (learn how to actually use it along the way). Estimated time: 4 hours. Performed by both.
4. Creating initial plots to attempt some visualization of correlation and better understanding data. Estimated time: 6 hours. Performed by N. Berzins.
5. Performing Time series analysis on the energy consumption pre-COVID to predict the energy consumption pattern if COVID pandemic didn't happen. Calculate delta between expected and real energy consumption. Estimated time: 8 hours. Performed by A. Rutmane.
6. Fit the events with Covid cases/dates. Estimated time: 4 hours. Performed by Anna Rutmane.
7. Perform regression analysis between delta(True v. Expected) Energy consumption and external factors (Covid cases and events). Estimated time: 6 hours. Performed by N. Berzins
8. Creating graphs/plots to showcase impact of COVID-19 on Energy consumption on each of the 3 Baltic countries. Estimated time: 4 hours. Performed by N. Berzins.
9. Create comparative plots between the countries. Estimated time: 5 hours. Performed by A. Rutmane.
10. Draw conclusions. Estimated time: 4 hours. Performed by both.
11. Find mistakes if any, redo the whole thing. Estimated time: 30 hours. Performed by both. (This is a real possibility, however can be ignored for now)
12. Create a poster for presentation. Estimated time 8 hours. Performed by both.