Data Science project D4: Why is electricity so expensive

https://github.com/Rikuklane/D4-Electricity-Prices

Richard Kuklane, Jakob Tamm, Georg Šumailov

Task 2 - Business understanding

Background

Our project is not made to specifically achieve financial gain for a specific business. It has rather exploratory and interest-driven motivation behind it. Therefore there will be a slight deviation in business understanding from the CRISP-DM process. However, it's very likely that its outcome might be financially beneficial to individuals or companies currently based in Estonia that pay for electricity they consume.

Background from an institutional perspective:

COVID-19 had a big impact on modern society as a whole because the routine and native lifestyle of many individuals was disrupted. For many of whom it meant spending more time at home and consuming more electricity as a result. Though on the other hand businesses went out of service.

This leaves for further discussion and speculation if in the context of Estonia COVID-19 had a negative or positive impact on electricity prices and consumption in general?

If any correlations would be found (for example lower energy consumption during lockdowns) then they would definitely be beneficial for energy producing companies to adjust and optimize energy production.

Business goal: Analyze COVID-19 impact on electricity prices and consumption in Estonia.

Business success criteria: Unfortunately business success criteria is subjective, assignee to deduct success of the project will be Jakob Tamm. It will be a success if any meaningful correlation will be found.

Background from an individual perspective:

With the duration of technology used by people in their homes increases, therefore increases electricity consumption and so the electricity bills. Combined with fluctuating electricity prices it might be useful for an individual or company to follow smart energy consumption theory in order to cut down expenses and save money.

Smart energy consumption implies that we know when to turn off or on our electricity devices, yet the best times in the context of Estonia remain unknown.

Business goal: Investigate how much money could be saved by applying smart consumption theory in Estonia.

Business success criteria: Electricity bill reduction by >7% (at the end of the day 7% could be anything, it's for the individuals to decide if in the end smart consumption theory is worth applying or not.

Situation assessment

Inventory of resources:

- **Software:**
    - For team collaboration: **Github**
    - For fetching data and overall structure of the project: **Python3**
    - For data exploration: **Jupyter Notebook**
    - For data visualization: **Python3 packages, e.g. matplotlib**
- **Data:**
    - **Nordpool datasets**
    - **Ilmateenistus datasets** (inquiry in process)
- **Contacts:**
    - **Alo Peets** (knows about everything related to the project)
    - **Richard Kuklane** (previous work experience with python)
    - **Jakob Tamm** (previous work experience with thinking)
    - **Georg Šumailov** (previous work experience.. :D)

Requirements, Assumptions, Constraints:

- **Requirements:**
    - Project must be easily replicated from start to end (pipeline, manual step-by-step tutorial)
    - Project must give an answer, result to the aforementioned business goals
    - Project output must be visualized in some comprehensible way

- **Assumptions**
    - Ilmateenistus 2021 dataset will be available soon
- **Constrains**
    - Project must be completed by 15.12.2021

Risks and contingencies

- If Ilmateenistus dataset for 2021 won't be given to us, then another source must be found

Terminology

- **Dataset** is a collection of related sets of information that is composed of separate elements but can be manipulated as a unit by a computer (Source)
- **Data mining** is the process of finding anomalies, patterns and correlations within large data sets to predict outcomes. Using a broad range of techniques, you can use this information to increase revenues, cut costs, improve customer relationships, reduce risks and more (Source)
- **Smart consumption** is interpreted in the context of this project as electricity consumption in a specific timeframe where electricity prices are generally lower compared to other timeframes
- **External parameter** is an electricity price influencing factor, for example wind intensity

Costs and benefits

- This project is of exploratory nature, therefore the possible costs and benefits from this project are highly subjective and speculative. For example if it is found that an individual residing in Estonia can cut expenses by 10% by applying smart consumption theory it is up to an individual if it is worth changing the contract with electricity provider or not. For some individuals 10% might be worth it and for some not. Same goes to companies.
- The cost both for companies or individuals would be either manually turning on and off electricity consuming devices at specified timeframes or acquiring energy price aware devices, which would turn on and off automatically depending on the energy price.

Data-mining goals

Data-mining goals:

- Report whether correlations between external parameters and electricity price in Estonia have been found

- Report on best general timeframe for energy consumption
- Report whether COVID-19 had possible impact on energy consumption and electricity prices in Estonia
- Data visualization for all aforementioned reports

Data-mining success criteria: Any meaningful correlations found conveyed by comprehensive data visualizations

Task 3 - Data understanding

Data gathering

Data requirements

- **Estonia 2016 - 2021 electricity prices** (nordpool, .xls(in reality html))
- **Estonia 2016 - 2021 electricity consumption** (nordpool, .xls(in reality html))
- **Estonia** (regional) **2016 - 2021 wind intensity** (ilmateenistus, .xlsx)
- **Estonia** (regional) **2016 - 2021 rain intensity** (ilmateenistus, .xlsx)
- **Estonia** (regional) **2016 - 2021 temperature** (ilmateenistus, .xlsx)
- **Estonia 2016 - 2021 natural gas price** (yfinance, pandas Dataframe)
- **Estonia** (regional) **2016 - 2021 CO2** (OurWorldInData, .xlsx / .csv / .xls / .json / .xml)
- **Estonia 2016 - 2021 solar intensity** (has not been found, .xlsx / .csv / .xls / .json / .xml)

Data availability: Currently we have the 2016-2021 data for electricity prices and electricity consumption from nordpool, 2004-2020 data for wind and rain intensity and temperature from ilmateenistus, 2018 - 2021 data for gas prices that are used by Eesti Gaas, 1830-2020 data for CO2 levels with yearly data only. Solar intensity seems to be nowhere to be found, easily available are only averages over the years or “climate norms”, possibility to write a letter to ilmateenistus and ask from there or to narrow the scope down a little bit.

Data selection criteria: From nordpool we will use three columns: date, hour, EE (estonia) for both electricity consumption and prices. From ilmateenistus we will use Tartu-Tõravere station information and we will use the next columns: year, month, day, hour, humidity, the amount of precipitation, air temperature, average wind speed. From yfinance we will use Dutch TTF Natural Gas Calendar and the columns Date and either Open or Close. From OurWorldInData we will use Annual CO2 emissions (per capita) and use columns year and CO2 emissions where Entity is Estonia.

Data description

Nordpool

- **Link:** https://www.nordpoolgroup.com/historical-market-data/
- **Format:** .xls file (though actually is in html format)
- **Issues:** some files have duplicate columns “EE”

Ilmateenistus

- **Link:** https://www.ilmateenistus.ee/kliima/ajaloolised-ilmaandmed/
- **Format:** .xlsx file
- **Issues:** 2021 data is missing, does not have solar intensity

YFinance

- **Link:** https://finance.yahoo.com/quote/TTF%3DF/
- **Format:** pandas Dataframe through yfinance python package
- **Issues:** The data only starts from October 2017, not sure if to use Close price or Open price, only daily format, no hourly movement.

OurWorldInData

- **Link:** https://ourworldindata.org/co2/country/estonia
- **Format:** .csv file
- **Issues:** 2021 data is missing, only has yearly data instead of hourly or daily.

Data exploration

Nordpool

Data was a bit difficult to read as it was in html format, though could be reformatted using python library BeautifulSoup4. The duplicate columns were easy to combine as if one had data then the other didn't and no row had data in both columns or in neither.

In data distribution there is one day each year in March when data is not measured. Also three times in five years there seems to be a 2-5 hour gap where there is no consumption (power loss?) Sometimes the electricity price is negative, though I know that it has actually happened during the recent years. Overall the data seems to be of good quality.

Ilmateenistus

Data is easy to read from the .xlsx file. Data is mostly intact, though out of the 149 040 values 24 000 values are empty regards to precipitation (years 2004-2012) and a maximum of 30 empty fields for other columns. Though starting from the year 2013 there are only 22 null value rows in the dataframe (wind intensity and air pressure) and starting from the year 2014 there are only 6 null values in the dataframe (wind intensity). Data seems to be of good quality starting from the year 2013 or 2014.

YFinance

Data is easy to read from the python package "yfinance". There are no null values and data seems to have correct values, though they are given by day, not by the hour. The prices are only available for the days that the stock market was open for trading.

OurWorldInData

Data is easy to read from the .csv file. Does not have any missing data and the yearly data that it has seems to be of good quality.

Data quality verification

Nordpool: The data is good enough to support the project goals, a small issue with the data was that sometimes correct data is displayed over multiple duplicated columns. Can provide electricity prices and consumption in Estonia.

Ilmateenistus: The data is good enough to support the project goals starting from years 2013 or 2014. Can provide humidity, rain and wind intensity, air pressure on sea level and temperature. The question remains if 1 weather station for the whole country is enough or multiple stations should be used.

YFinance: This data is good enough to support the project goals for finding correlations. Though the data has past 4 years of data for workdays only (when could trade on stock market) and only daily data.

OurWorldInData: This data is not good enough to support our project goals and will not be used as it will be really hard to see if it affects the prices of electricity.

Task 4 - planning your project

Our project will consist of 5 main tasks. The first one is currently in the endphase and it's actually finding the necessary datasets for the analysis. This mainly includes searching the internet for weather data and data about electrical prices. We've also written e-mails to ask for further data to make our analysis more versatile. This task has taken us all around 3-4 hours.

The next task is actually collecting and downloading the raw data. Since we've located the datasets already, this part shouldn't take longer than 15-30 minutes each.

The third task is processing the data (checking for missing values/irregularities) and exploring/understanding it. This is the part where we also get rid of all the data we don't deem necessary for our analysis. We will be splitting this task and it should take 3-6 hours each.

The fourth and most important task is performing the analysis on the gathered data. This involves finding patterns in the weather and electrical prices and comparing them with the coming of Covid-19 etc. Since this is the backbone of the entire project, this will take us the biggest amount of time - 15-20 hours each.

The fifth task is making our analysis presentable. At this point we should have an excellent understanding of our data, but this needs to also be conveyed to the listeners, which means visualising the data and wording the results of our analysis in a coherent way. This part also includes preparing our presentation for others. This task will take around 4 hours each.

Currently we are using and planning to use GitHub, Jupyter Notebook, Python (pandas, bs4, plotly, etc..) and Conda. This list might become longer as the need for some certain tools will become apparent in the later stages of this project.