Final project documentation - PDA

Contents

Project plan

Context

Domain knowledge and the business context *Briefly describe the business/organisation and where your analysis fits within its aims/activity.*

For the purposes of this project, I am imagining I am analysing data for the Energy Saving Trust in Scotland. The Energy Saving Trust aims to provide data-driven guidance to government, businesses and household energy consumers to support these stakeholders to do their part in helping to reduce non-renewable energy use and switching to renewables.

In Scotland, according to data from 2022, 41.2% of electricity is consumed in domestic households, and the majority of electricity used (overall) comes from renewable sources (Scottish Government: Energy Statistics for Scotland - Q3 2022).

Here, the (imaginary) brief is to understand how to predict domestic electricity consumption according to time and weather, so as to understand how to make further reductions in energy use and/or ensure electricity demand can be met with renewable sources.

Without having access to sufficiently granular data for domestic energy consumption in Scotland, I am using existing historic data from London households (see below).

Business intelligence and data-driven decision making *What insights can the business/organisation gain from your analysis and how will your analysis help the business/organisation make better decisions?*

Understanding which features are good predictors for household energy use enables the business to:

- model energy consumption in upcoming years (and seasons) with different weather scenarios (i.e. as climate changes)
- predict times of high and low energy demand (low demand is important, because renewable sources are often turned down at this time to manage load in the grid)
- calculate potential energy and cost savings from any improvements to domestic energy efficiency

There may also be ways to cluster households by their energy consumption patterns, which may be a useful way to segment domestic energy customers and provide more targeted information and policy recommendations.

Data

Internal and external data sources The data sources for this project are all external:

1. **London Energy Data** - a single csv file with daily aggregated energy usage for 5,567 London households from November 2011 to February 2014 from the UK Power Networks' SmartMeter Energy Consumption Data in London Households project, provided to the public domain (CC0) by Emmanuel F. Werr on Kaggle. URL: https://www.kaggle.com/datasets/emmanuelfwerr/london-homes-energy-data. (downloaded on August 19, 2023). This csv file contains 3 attributes (household id, date, total energy consumption (in kilowatt-hours, kWh)) and 3510433 observations, where one observation is one household's energy consumption on that date.
2. **London Weather Data** - a single csv file with daily historic weather observations in London from 1978 to 2021 sourced from the European Climate Assessment & Dataset (ECAD), provided to the public domain (CC0) by Emmanuel F. Werr on Kaggle. URL: https://www.kaggle.com/datasets/emmanuelfwerr/london-weather-data (downloaded on August 19, 2023). This csv file contains 10 attributes (date, and nine weather measurements, including min, max and mean temperature, number of hours of sunshine and total precipitation) and 15341 observations, where one observation is one day's weather. The weather measurements are a mix of totals, averages, minimums and maximums across the different measures, although it is not clear for some values whether they are totals or averages, since the original element code and the cleaning script is not included.
3. **UK Power Networks' original data** - original data from the UK Power Networks' project mentioned above, provided as half-hourly measurements of household energy usage for 5,567 London households, including tariff types, and covering November 2011 to February 2014. URL: https://data.london.gov.uk/dataset/smartmeter-energy-use-data-in-london-households (zip file of low-carbon-london-data-168-files (758.86 MB) downloaded on August 19, 2023) and reused here under the CC-BY license.

The original research project by UK Power Network included two groups of customers: one group (~1100 households) had dynamic energy pricing, and received warnings about higher and lower prices for the next day; the other set (~4500 households) had fixed prices. The project aimed to understand how pricing warnings might affect energy consumption behaviour.The daily aggregated London energy dataset provided on Kaggle does not include this information about the tariff settings and warnings, therefore there is a missing variable here. However the aggregated data is smaller and easier to work with than the original research data provided by the UK Power Network (which is 167 million rows in total, with half-hourly measurements). Given the timeframe, I have chosen to work with the aggregated data (at least initially).

More information about the original Low Carbon London project, including final reports, is available from https://innovation.ukpowernetworks.co.uk/projects/low-carbon-london/.

Types of data *What kind of data did you work with? E.g. categorical and numerical data and their sub-types.*

- The **london_energy** dataset has 3,510,433 rows and 3 columns:
  - 1 character column, which is the categorical household identifier (lc_lid)
  - 1 numeric, which is the continuous numeric energy usage (kwh)
  - 1 date, which is the date of the energy usage (date)
- The **london_weather** dataset has 15,341 rows with 10 columns, all numeric
  - date - recorded date of measurement - (dbl)
  - cloud_cover - cloud cover measurement in oktas - (dbl)
  - sunshine - sunshine measurement in hours (hrs) - (dbl)
  - global_radiation - irradiance measurement in Watt per square meter (W/m2) - (dbl)
  - max_temp - maximum temperature recorded in degrees Celsius (°C) - (dbl)
  - mean_temp - mean temperature in degrees Celsius (°C) - (dbl)
  - min_temp - minimum temperature recorded in degrees Celsius (°C) - (dbl)
  - precipitation - precipitation measurement in millimeters (mm) - (dbl)
  - pressure - pressure measurement in Pascals (Pa) - (dbl)
  - snow_depth - snow depth measurement in centimeters (cm) - (dbl)]

- The **original research data** for this project was downloaded as a zip file, with the data split into 167 individual .csvs, each with up to 1 million rows (to a total of ~167 million observations), and all with 4 columns:
  - 2 categorical character variables: lc_lid (household id, same as above), stdorToU (the pricing tariff group the household was assigned to)
  - 1 continuous numeric variable: KWh/hh (kilowatt-hours of energy used per half-hour)
  - 1 datetime (dttm) variable: DateTime (the date and time that the half-hour measurement was recorded)

I did not work further with this third data source, due to time and processing limits.

Data formats *What format did your data come in? E.g. all downloaded flat files (CSV) or any data from APIs, scraping etc.*

All data were downloaded as CSV files from source websites.

Data quality and bias *Briefly describe the quality of the data and whether you have any reasons to suggest the data is biased e.g. only data from a specific demographic even though a broader demographic would be of interest to the organisation.*

According to the original research project, "the customers in the trial were recruited as a balanced sample representative of the Greater London population" (]data.london.gov.uk](https://data.london.gov.uk/dataset/smartmeter-energy-use-data-in-london-households)).

Energy Saving Trust work across the UK, and therefore we should take care about extending any conclusions from this data beyond London, especially where demographics and energy needs may differ substantially.

Furthermore, we know the original research project split households into two different pricing tariffs, and their experiment was designed to test how/whether the pricing tariff made a difference to energy usage. We do not have these tariff groups in our data, which means we may be missing the impact of a predictor variable (i.e. there is a hidden confounding variable in our data). It would be better to understand the tariff group for each household from the original data, and bring this into our daily London energy data - I haven't done this for this project due to time and processing constraints.

Ethics

Ethical issues in data sourcing and extraction *Do you have any ethical concerns regarding the sourcing and extraction of your data?*

- Primary data source: for the household energy consumption data, the data were collected as part of a research project, working with EDF customers with their consent (noted in an early progress report: https://innovation.ukpowernetworks.co.uk/wp-content/uploads/2019/05/Six-Monthly-Project-Progress-Report-June-2011.pdf). The data used does not contain any information about the household itself, other than an identifier value, so we do not know location or any personal information about the inhabitants.
- Secondary data: data were sourced from Kaggle, for which the author had processed the data from the primary sources. The processing scripts and details were not provided, therefore I had to rely on the author doing this correctly. Some assumptions have been made as to what the weather data variables mean, given the original source had multiple options for each of the variables inferred from the column names. However, I do not think there are any ethical concerns about reusing this data - while each household's energy usage behaviour can be visualised, it would be very difficult to infer any personal information from this or misuse this data in any way.

Ethical implications of business requirements *Are there any ethical implications of the business requirements?*

Categorising households according to their energy consumption patterns may reveal personal information about how a house is used, including whether it is vacant (many days with low/no energy usage), or could be used to infer properties of the household (size, insulation, whether there are energy intensive devices being run). So the business requirements are not free from ethical implications, and we should consider how the resulting data or model may be used by the business or any others who see any public outputs. For example, burglars could target households with very low usage on predicted zero-energy days, with the assumption they may be unoccupied residences. However, I consider this unlikely, and overall I think the probability of any severe implications from misuse is low.

Analysis

Stages in the data analysis process *What were the main stages in your data analysis process?*

1. **Exploration of data** - to:

a. understand what's possible, with which data sources - see brief data exploration notebook
b. analyse the data overall: summary statistics, correlations, patterns, working out how to clean/wrangle/transform data for clustering and what kinds of clustering splits to look out for - see notebooks for data exploration, weather data exploration, household energy stats and feature engineering for clustering (including first attempt at K-means clustering)

2. **Cleaning, preparing, joining data** - informed by the above exploration, see scripts to prepare household_energy_data (from London energy dataset), prepare weather data (from London weather dataset) and join the two datasets together.

a. Steps for cleaning household energy data included (i) clean column names, (ii) trim dataframe dates to remove low-number early dates and unusual end date values and (iii - optional) remove some households due to insufficient data and/or high percentage 0 kWh days (the final joined data used in clustering did not include this data removal step, it included all households' data at first and households with insufficient data were removed during the clustering preparation process, see below)
b. Steps for cleaning weather data included (i) trim to date range matching the energy data, (ii) factorise some weather observations (e.g. snow to TRUE/FALSE) nad (iii) add some data quality indicators where temperature variables do not make sense (e.g. min_temp > mean_temp for that day)

3. **Perform K-means clustering** - see K-means clustering notebook, including data preparation for clustering, analysis of clustering model performance and visualisation of the resulting clusters
4. **Create visualisations to present** to the "client" (in a live 10-minute presentation to CodeClan instructors and fellow students, August 30th) - see presentation graphs notebook
5. **Create presentation** - see PDF of slides

Tools for data analysis *What were the main tools you used for your analysis?*

All analysis was conducted in R (version 4.3.1) using RStudio.

The analysis notebooks used the following R packages:

- tidyverse - for wrangling and analysing
- lubridate - to work with datetime data
- ggplot2 - to visualise data and produce graphs to present
- tsibble - to work with time series data
- slider - to make a rolling average of energy usage, to show a smoother line on the plot
- GGally - for ggpairs to explore correlations
- psych - for pairplots to explore correlations
- cluster - for k-means clustering
- corrplot - for correlation plot
- broom - for k means optimisation stats
- ggsignif - for silhouette method
- rstatix - for silhouette method

- factoextra - for silhouette method
- feasts - to try time series forecasting (deprioritised in this project)

The presentation was made using Google Slides.

Descriptive, diagnostic, predictive and prescriptive analysis Please report under which of the below categories your analysis falls **and why** (can be more than one)

```
**Descriptive Analytics** tells you what happened in the past.

**Diagnostic Analytics** helps you understand why something happened in the past.

**Predictive Analytics** predicts what is most likely to happen in the future.

**Prescriptive Analytics** recommends actions you can take to affect those outcomes.
```

This analysis comprises **descriptive** and **diagnostic** analytics - I explore historic energy usage and use clustering and correlations to identify potential variables that may explain energy usage (here: season, weekday type).

PDA Outcomes

Working with Data (J4Y6 35)

1. Plan an analysis to provide business intelligence

- 1.1 Business intelligence and data-driven decision making
- 1.2 Domain knowledge and the business context
- 1.4 Internal and external data sources
- 1.5 Data quality
- 1.6 Stages in the data analysis process
- 1.7 Descriptive, diagnostic, predictive and prescriptive analysis
- 1.9 Ethical implications of business requirements
- 1.10 Tools for data analysis

2. Extract data from a variety of sources

- 2.1 Tools for querying data sources
- 2.2 Types of data (categorical and numerical data and their sub-types)
- 2.3 Data formats
- 2.6 Data quality including data bias
- 2.7 Ethical issues in data sourcing and extraction

4. Analyse data to provide business intelligence

- 4.7 Role of domain knowledge in interpreting analyses