Dataset Documentation

Machine Learning for Efficient Data Management and Sustainability in UK's Energy Landscape: An ESG Perspective based on Geographic Information Systems

November 15, 2023

1 Overview

This document provides documentation for six datasets related to energy usage and sustainability in the UK:

- Dataset 1: Indicative power emissions pathway to 2037. This dataset provides an indicative pathway for power emissions in the UK from 2023 to 2037. It is based on projections from the UK government's Net Zero Strategy.
- Dataset 2: Reducing emissions across the economy. The dataset is about reducing emissions across the economy in the UK. It discusses the UK's goal of achieving net zero emissions by 2050. The article also details specific steps the UK is taking to reduce emissions in the power, transport, and industry sectors. Some of the important points from this article are that the UK plans to have a fully decarbonised power system by 2035 and that the UK is investing heavily in electric vehicles.
- Dataset 3: Gas section 4 energy trends. The dataset is about the UK's gas sector. It includes data on upstream production, trade, and demand. The data is quarterly and monthly, and the most recent data is from September 2023. The dataset also includes data on gas prices. Gas prices have been rising in recent months, due to a number of factors, including the war in Ukraine and the global economic recovery.
- Dataset 4: Postcode-level gas statistics 2020 (experimental). The dataset is about domestic gas consumption in the UK. It discusses gas consumption estimates at the postcode level for 2020. The data is provided by the Department for Business, Energy and Industrial Strategy.
- Dataset 5: ESG footprint of Synthetic UK Individuals and Businesses. The dataset contains information on 1000 synthetic UK individuals and businesses, including their borough or county, primary sector, entity trade name, annual turnover, entity status, country of primary operation, and energy consumption data. The energy consumption data is broken down by fuel type (coal, electricity, gas, natural gas, and petrol) and sector (toe). The dataset also includes geographic codes, local authority names, SIC group codes, section codes, sector codes, and structure type codes.
- Dataset 6: UK Power plants. The dataset containing information about renewable energy power plants in the United Kingdom. The dataset is divided into separate CSV files for each country, and the data availability and accuracy varies from country to country.

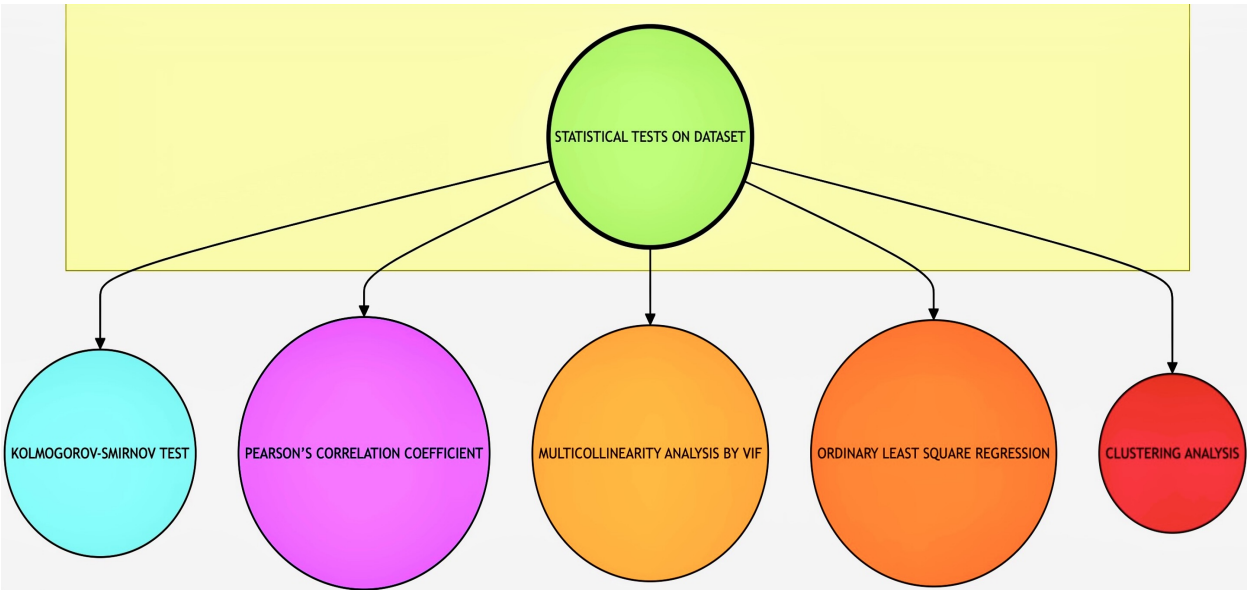

Figure 1: List of statistical tests on dataset for validation

1.1 Statistical Tests On Dataset

Understand Data Characteristics The List of statistical test on dataset for validation figure can be used to understand the overall characteristics of the dataset. Basically it helpes to know the pattern of data in which variables have been tested for normality and multicollinearity, and which variables are correlated with each other. This is the significance of the dataset.

Identify Anomalies The Pearson's Correlation Coefficients Between Variables figure can be used to identify anomalies in the data (in figure 2 and 3). For example, if reseacher see a correlation coefficient that is very high or very low, this may indicate that there is an anomaly in the data. Researcher can then investigate this anomaly further to see if it is due to a data entry error or a genuine outlier.

Multicollinearity Test The Multicollinearity Analysis by Variance Inflation Factor (VIF) figure can also be used to identify anomalies in the data. For example, if the researcher see a VIF value that is greater than 5, this indicates that there is a high degree of multicollinearity between the variables. This can lead to problems with the regression model, so one may need to remove one or more of the variables from the model.

Form of Hypotheses The Pearson's Correlation Coefficients Between Variables figure can be used to form hypotheses about the relationships between the variables in the dataset (as shown in figure 9). For example, if researcher see a strong positive correlation between two variables, one may hypothesize that these variables have a causal relationship.

Uncover Patterns The Clustering Analysis IQR (Interquartile Range) method figure can be used to uncover patterns in the data. For example, researcher can see which data points cluster together and which data points are outliers.

Inform Further Analysis The results of the statistical tests and the insights gained from the figures can be used to inform further analysis of the dataset (as shown in figure 5). For example, if researcher identify any anomalies in the data, one may need to remove them before further analysis. If researcher identify any strong correlations between variables, one may want to test these correlations using a regression model. And if one identify any patterns in

Test	Statistic	Value
One-Sample Kolmogorov-Smirnov Test for Coal Consumption	D-statistic	0.48300000000000004
	p-value	3.1227482537563926e-215
One-Sample Kolmogorov-Smirnov Test for Electricity Consumption	Kolmogorov-Smirnov statistic	0.3941
	p-value	0.0000
One-Sample Kolmogorov-Smirnov Test for Gas Consumption	Kolmogorov-Smirnov statistic	0.3434
	p-value	0.0000
One-Sample Kolmogorov-Smirnov Test for NG Consumption	Kolmogorov-Smirnov statistic	0.3631
	p-value	0.0000
Hypothesis test: Anova	P-value	**0.3948 (not significant)**
Significance level (alpha)		0.05

the data, the researcher may want to investigate these patterns further using clustering or other machine learning algorithms.

Prepare Data for Modeling Insights derived from the statistical tests and visualizations are instrumental in readying the data for modeling purposes. For instance, should the data exhibit non-normal distribution, it necessitates transformation prior to its application in regression modeling. This step ensures the data's suitability for accurate and effective analysis.

1.2 Dataset type

The dataset type is tabular, and dataset format is CSV.

2 Purpose and Application of Datasets

The datasets are designed to serve several critical purposes, including:

- Evaluating the United Kingdom's advancement towards achieving its net zero emissions goal.
- Pinpointing sectors and activities where emission reductions are most achievable.
- Crafting policy frameworks and action plans aimed at minimizing power-related emissions.
- Analyzing temporal trends in energy consumption.
- Uncovering primary factors contributing to energy consumption.

- Gauging the effectiveness of various environmental sustainability initiatives.
- Constructing predictive models to forecast future patterns in energy use.

3 Distribution

All six datasets are publicly available and can be downloaded from the UK government website and Kaggle. The below dataset link have been given below:

References

[1] UK government. Net Zero Strategy. 2021. https://www.gov.uk/government/publications/net-zero-strategy

[2] UK government. Gas section 4 energy trends. 2023. https://www.gov.uk/government/statistics/gas-section-4-energy-trends

[3] UK government. Postcode-level gas statistics 2020 (experimental)2023. https://www.gov.uk/government/statistics/postcode-level-gas-statistics-2020-experimental

[4] OpenNetZero. UK renewable power plants dataset. https://opennetzero.org/dataset/uk-renewable-power-plants?q=Renewable+Energy+Capacity+and+Potential+in+the+UK+shapefile

[5] Nayaone. ESG Synthetic UK Population and Businesses. https://www.kaggle.com/datasets/nayaone/esg-synthetic-uk-population-and-businesses