DATA MINING

Umar Farooq 22i1978/ Anas Hamid 22i1951

Energy Demand Analysis Report

Project Title: Energy Demand Clustering and Forecasting
Data Range: Jan 1, 2018 – May 19, 2020
Observations: 10,000
Features Used: 43 original columns; 7 selected for clustering

1. Data Overview

- **Source:** Historical energy usage and weather conditions.
- **Variables include:**
  - Time-based: hour, day_of_week, month, season, etc.
  - Demand: demand, temp_demand_ratio, demand_zscore, etc.
  - Weather: temperature, humidity, wind speed, pressure, etc.
- **Normalized & Z-Scored** values used for clustering.

2. Data Preprocessing Summary

- **Features selected for clustering:**
  demand, temperature, humidity, windSpeed, hour, day_of_week, month
- **Techniques used:**
  - Z-Score Normalization
  - PCA for dimensionality reduction
  - t-SNE for visualization
- **Dimensionality Reduction:**
  - **PCA explained variance:** PC1 = 19.95%, PC2 = 18.64%
  - Top influencing features: temperature, demand, humidity, hour

3. Clustering Algorithms Evaluation

A. K-Means Clustering

- **Optimal K:** 6 (based on Silhouette Score and Elbow method)
- **Silhouette Score:** 0.1435
- **Most interpretable clustering**

B. DBSCAN

- **Clusters found:** 14
- **Silhouette Score (excluding noise):** 0.2411
- **Noise points:** 94.85% of data → too sparse

C. Hierarchical Clustering

- **Chosen Clusters:** 5
- **Silhouette Score:** 0.0771
- **Result:** Less effective than K-Means or DBSCAN

4. K-Means Cluster Interpretation

Cluster	Description	Count	Avg Demand	Avg Temp	Common Hour	Common Day	Correlation
0	Medium Demand - Moderate Morning	2333	5685.28	59.11°F	15	Thursday	-0.042
1	Medium Demand - Moderate Morning	1904	6696.71	60.53°F	11	Tuesday	0.137
2	High Demand - Hot Afternoon	1849	13829.33	73.11°F	18	Wednesday	0.242
3	Medium Demand - Moderate Morning	880	9504.53	63.36°F	00	Thursday	0.297
4	Medium Demand - Hot Morning	1228	7478.65	79.25°F	02	Friday	0.168
5	Medium Demand - Moderate Afternoon	1806	6427.68	58.48°F	19	Sunday	0.102

- **Interpretability Scores:**
  - Highest: Cluster 2 – 81.05/100 (Excellent)
  - Lowest: Cluster 4 – 61.58/100 (Good)

5. Key Findings from Clustering

- Clear temporal and temperature-based demand groupings.
- High demand mostly aligns with high temperature and afternoon time blocks.
- Weekdays show distinct demand profiles.
- These clusters can drive **demand-side management**, **forecast model segmentation**, and **policy planning**.

6. Forecasting Model Preparation (Upcoming)

To be executed on the clustered data (or overall):

💡 Algorithms Planned:

- **Linear Models:**
  - Linear Regression
  - Ridge Regression
- **Ensemble Models:**
  - RandomForestRegressor
  - GradientBoostingRegressor
  - XGBoost
- **Time Series Models:**
  - ARIMA
  - SARIMAX
- **Deep Learning:**
  - LSTM (Keras Sequential Model)

🔁 Evaluation Metrics:

- MAE (Mean Absolute Error)
- RMSE (Root Mean Squared Error)
- R^2 Score (Coefficient of Determination)
- MAPE (Mean Absolute Percentage Error)
- Forecast Plots (Actual vs Predicted)

7. Exports & Files Generated

- energy_data_with_clusters.csv — clustered data with labels
- clustering_report_summary.csv — interpretability and summary
- Visualizations: PCA scatter plots, t-SNE, dendrogram, elbow curve, silhouette plots

8. Recommendations & Next Steps

1. **Use Clusters for Forecasting:**
   Train separate models per cluster to capture seasonality and behavior.
2. **Peak Demand Handling:**
   Focus on Cluster 2 for alert systems and infrastructure planning.
3. **Incorporate External Factors:**
   Test weather and holidays as additional regressors.
4. **Model Comparison:**
   Evaluate all forecasting models and choose based on metrics + interpretability.