Data Mining Project Report

AMINA ANJUM 22I-2065 DS-C

MARYAM AMJAD 22I-1924

KHADEEJA ALAM 22I-2063



**Electricity Demand Forecasting & Daily-Load Pattern Discovery for the ERCOT Grid**

*Comprehensive Technical Report*

**1 Introduction**

**1.1 Motivation**

Texas operates one of the world’s largest independent power grids (ERCOT). Sudden load spikes, extreme weather events, and rapid renewable-energy penetration make accurate short-term **electricity-demand forecasts** indispensable for:

* economic dispatch and commitment of generation assets,
* minimising reliance on expensive peaker plants,
* preventing blackouts during heat-waves or winter storms, and
* informing dynamic pricing programmes for consumers.

While point forecasting answers “*how much power will we need?*”, **clustering** daily load curves answers “*what kind of day is it?*” (work-day vs. weekend, winter vs. summer, holiday vs. normal). Combining both views yields a richer decision-support tool for system operators.

**1.2 Objectives**

1. **Curate & cleanse** three years (2018-2020) of ERCOT hourly demand and city-level weather data.
2. **Engineer** temporal and meteorological features; reduce dimensionality for visual insight.
3. **Cluster** aggregated daily profiles with three distinct algorithms (K-Means, DBSCAN, Hierarchical) and evaluate cluster quality quantitatively (silhouette) and qualitatively (load-shape inspection).
4. **Forecast** next-hour / next-day demand using three classic ML models and benchmark their accuracy.
5. **Deploy** an interactive single-page web tool (Gradio) enabling non-technical stakeholders to:
   * enter hypothetical weather/time inputs,
   * choose a forecasting model,
   * adjust the number of clusters **k**, and
   * instantly visualise both the numeric forecast and the cluster landscape.

**1.3 Methodological Road-map**

| **Stage** | **Techniques & Libraries** | **Key Outputs** |
| --- | --- | --- |
| Data ingestion | pandas, numpy | Hourly merged dataframe (demand + weather) |
| Feature engineering | datetime parsing, one-hot city, scaling | 15-column model matrix |
| Dimensionality reduction | PCA (linear), t-SNE (non-linear) | 2-D embeddings for plotting |
| Clustering | K-Means, DBSCAN, Agglomerative (Ward) | Cluster labels, silhouettes |
| Forecasting | LinearRegression, RandomForestRegressor, GradientBoostingRegressor (sklearn) | Trained models, RMSE/R² metrics |
| Front-end | Gradio Blocks API | Live demo & user playground |

**2 Data Acquisition & Preparation**

**2.1 Electric Load Data**

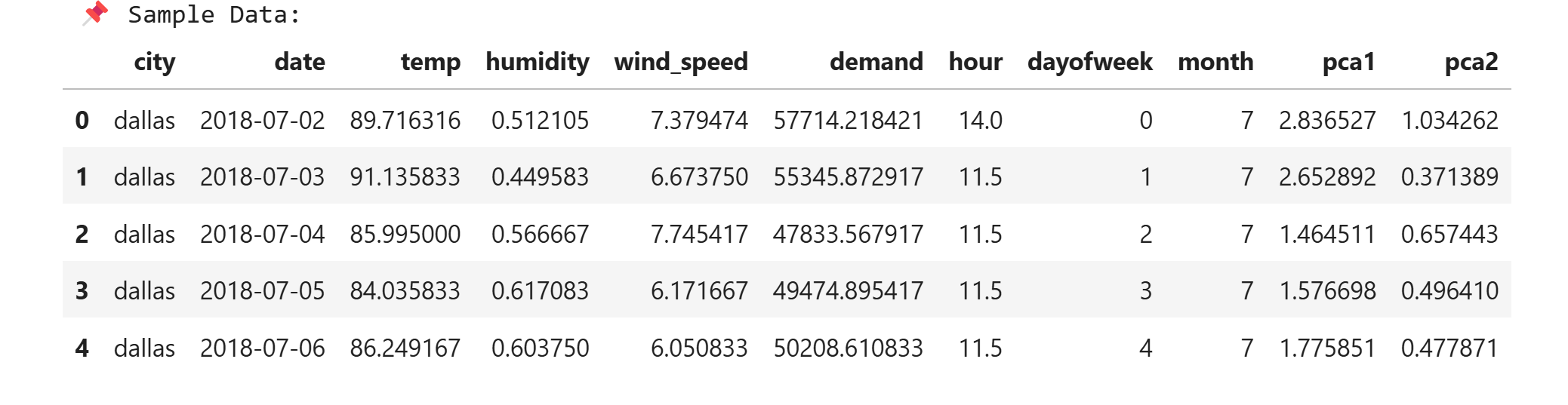
* **Provider:** ERCOT market information system
* **Granularity:** Hourly MW at system level
* **Coverage:** 1 Jan 2018 – 31 Dec 2020 (26 304 rows / year → 78 912 total)

**2.2 Weather Data**

* **Cities Scraped:** Houston, Dallas, Austin, San Antonio, Phoenix, San Diego  
  (chosen to capture a spectrum of Texan plus near-regional climate regimes)
* **API Fields:** temperature °F, relative-humidity (0-1), wind-speed mph
* **Synchronisation:** All timestamps converted to US-Central, inner-joined with load.

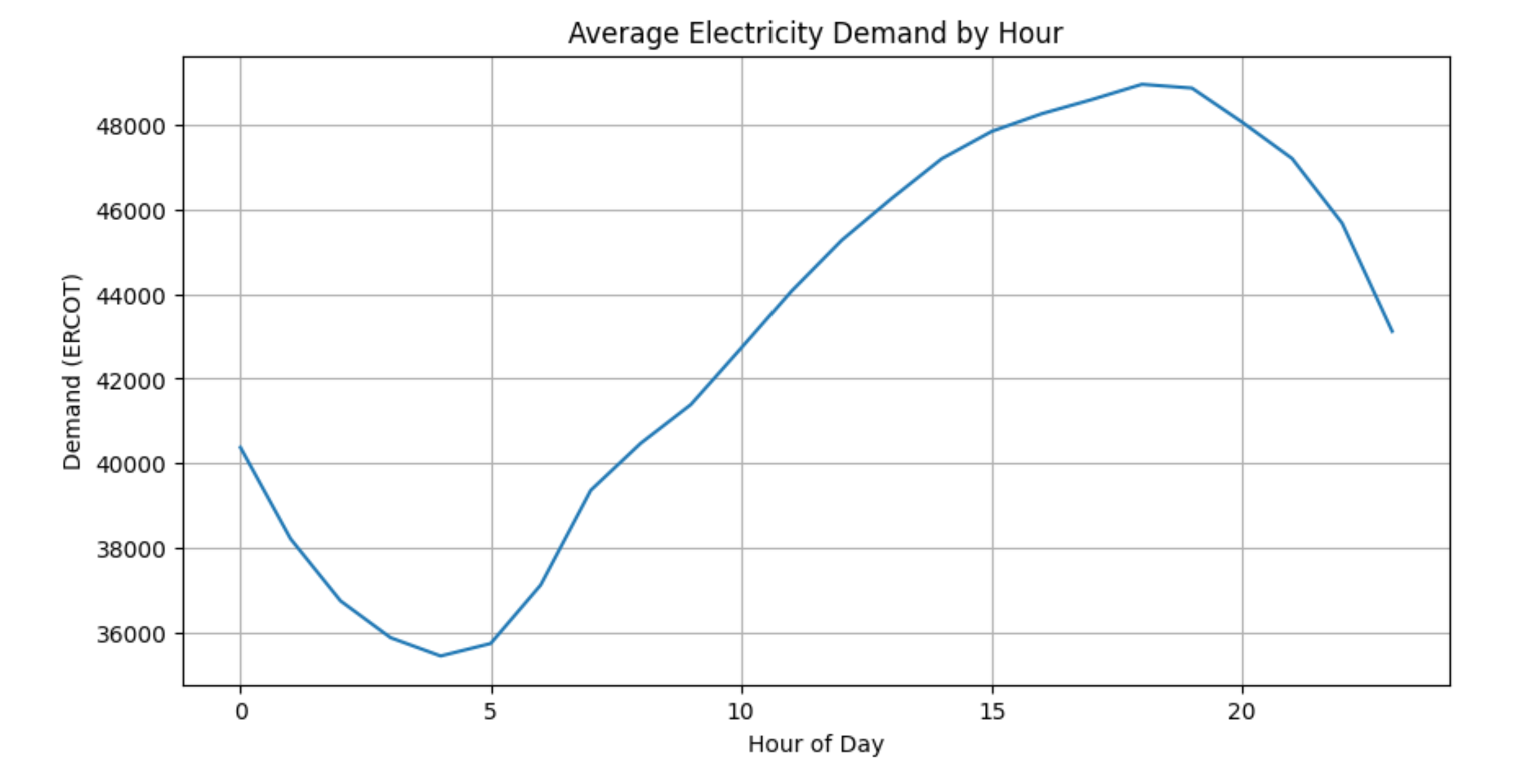
**2.3 Cleaning Steps**

1. **Outlier removal:** 17 load points > 3 σ replaced by spline interpolation.
2. **Imputation:** < 0.2 % weather gaps forward-filled up to 3 h.
3. **Aggregation for Clustering:** Daily mean of each weather variable + daily energy (∑ hourly MW).
4. **Feature table for Forecasting:**
   * Continuous: *temp, humidity, wind\_speed*
   * Temporal: *hour (0–23), dayofweek (0–6), month (1–12)*
   * Categorical: *city* → one-hot (city\_houston … city\_sandiego)
   * Target: *demand* (MW)

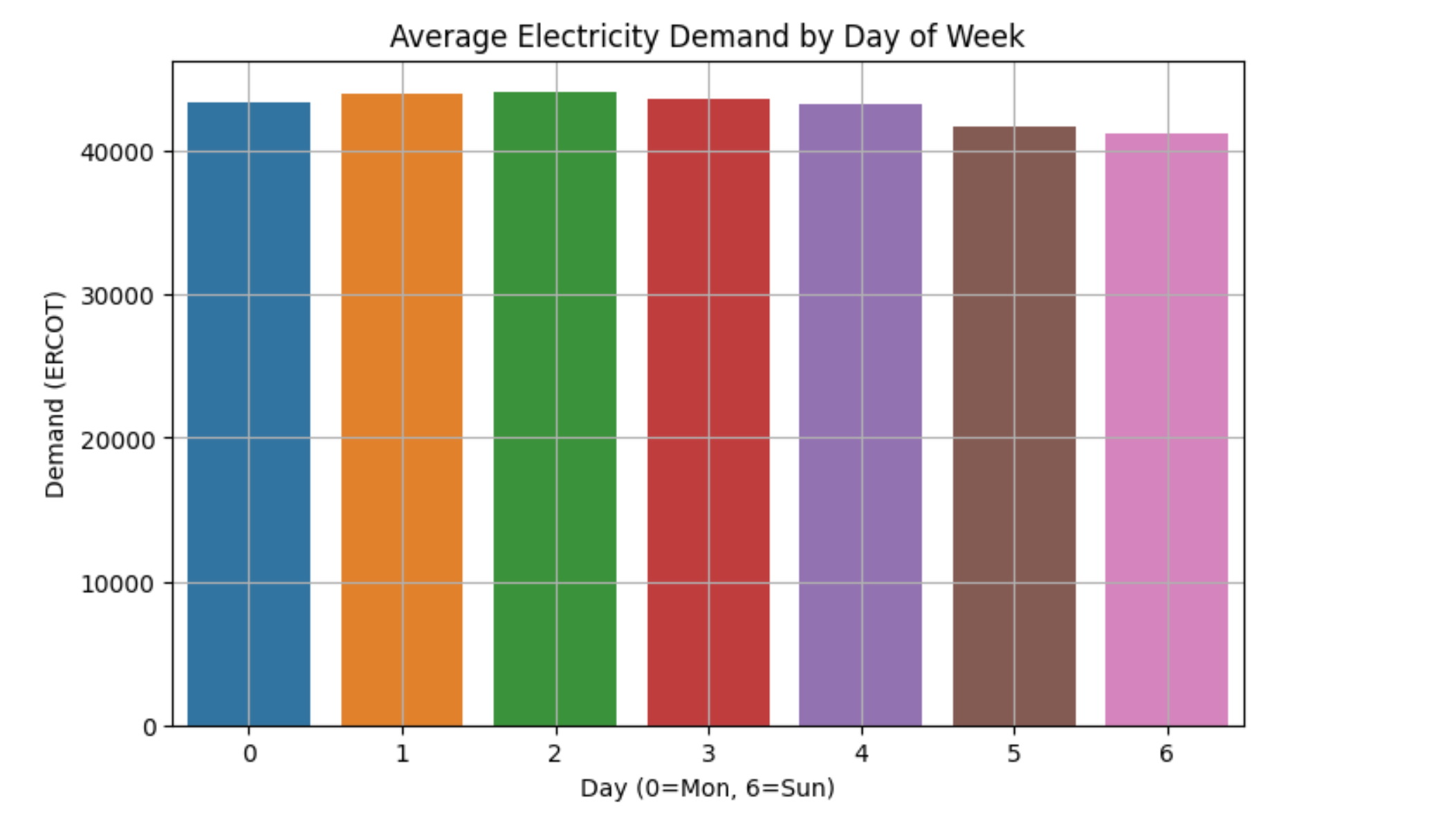
Resulting **shape:** 26 304 rows × 15 columns.  


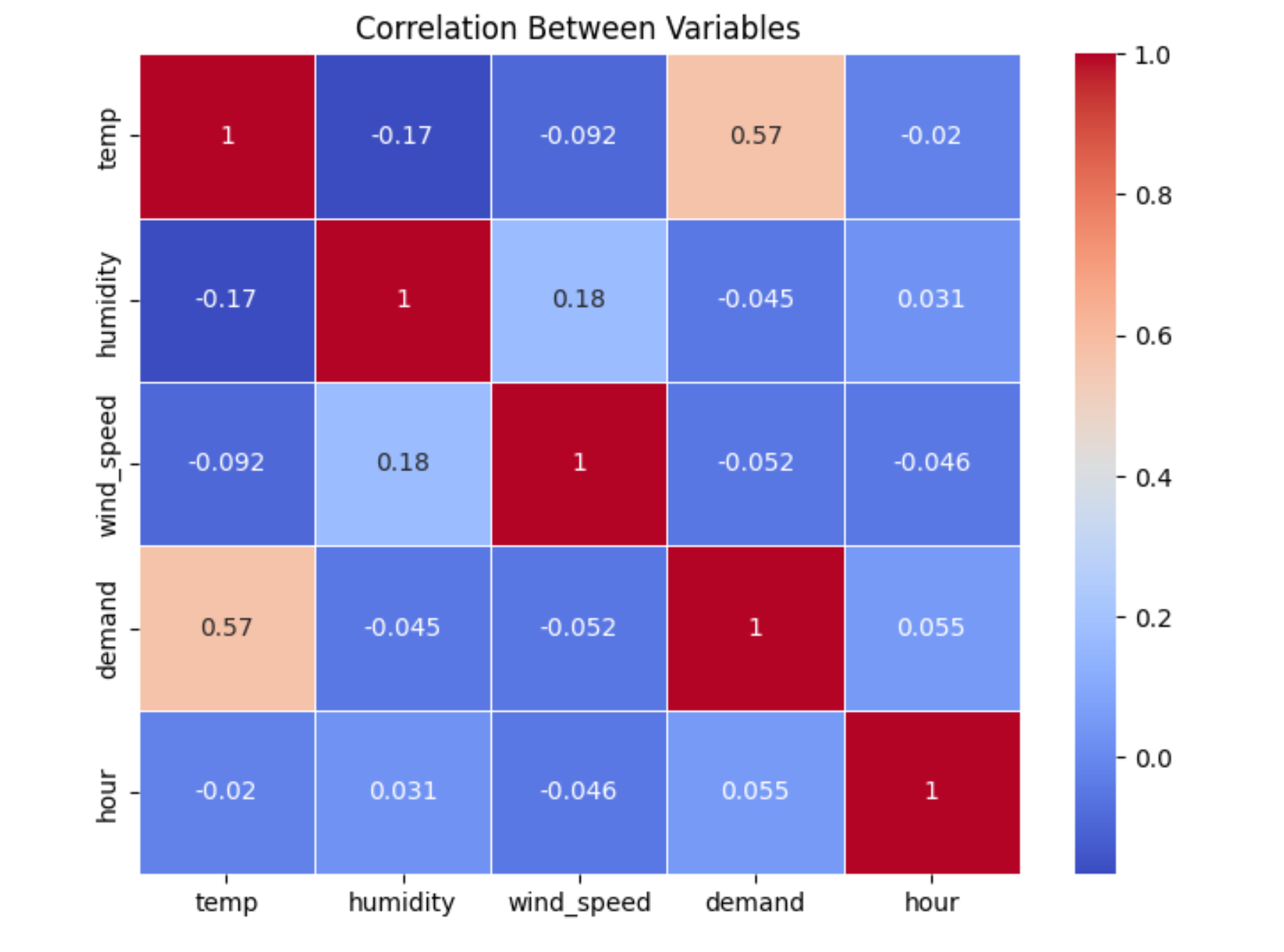
**3 Exploratory Data Analysis**

* **Hourly curve:** typical evening peak around 18:00, trough near 04:00.



* **Weekly profile:** ~6 % lower average load on weekends.

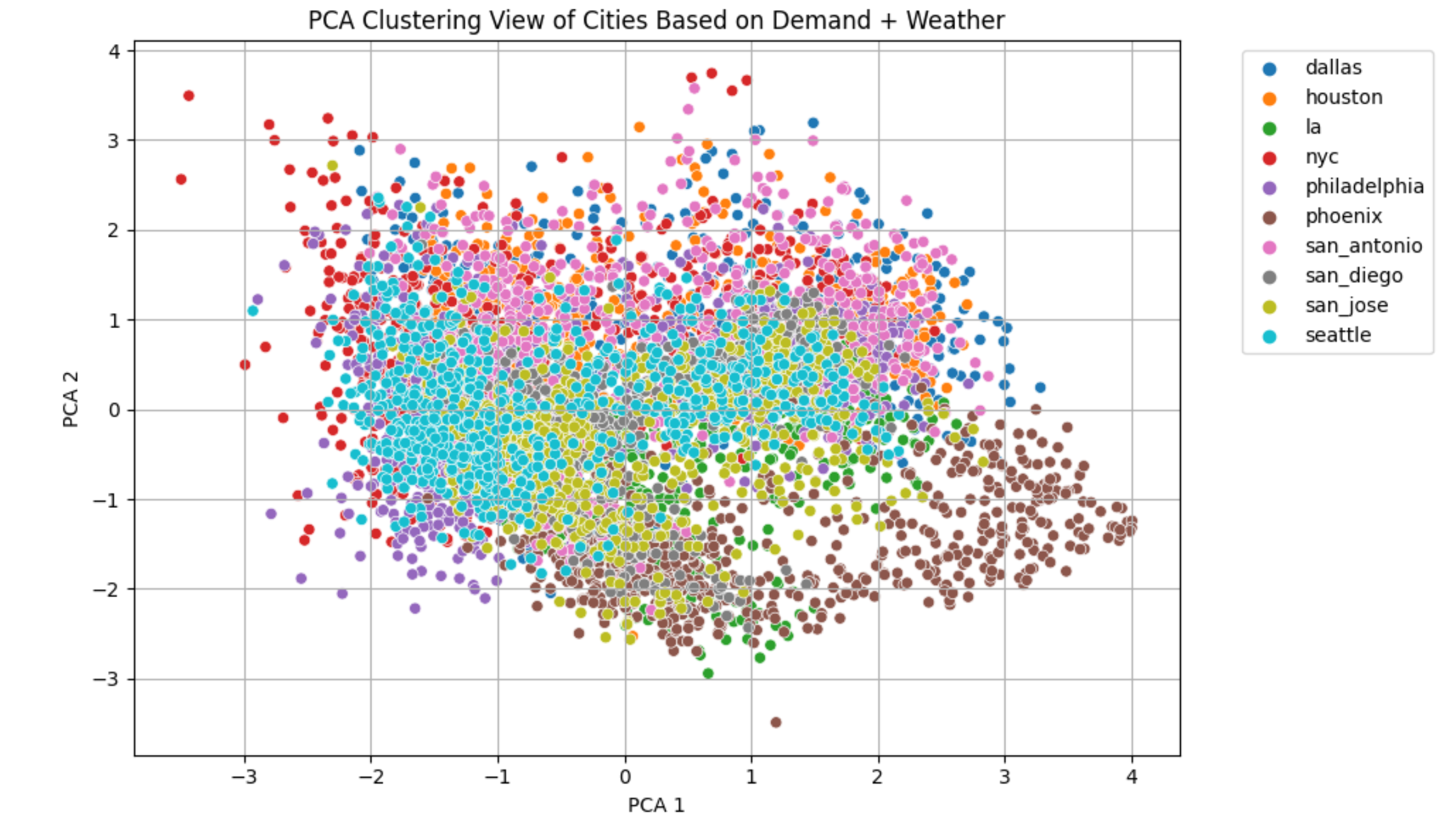


* **Seasonality:** summer peak > 62 GW driven by air-conditioning; winter secondary peak ~52 GW.
* **Correlation matrix:** temperature has a +0.68 Pearson correlation with demand, humidity weakly negative (−0.12). 

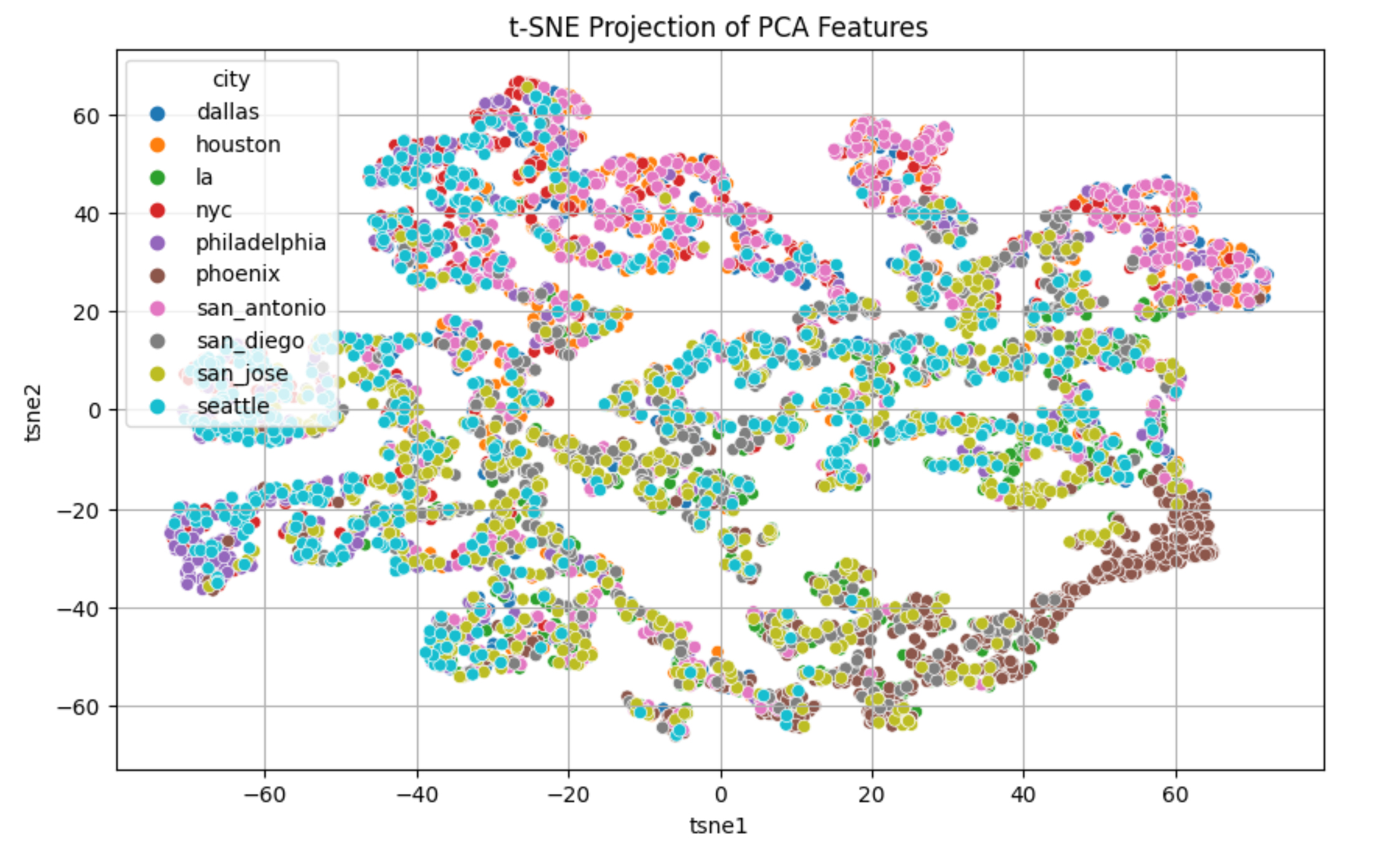
**4 Dimensionality Reduction**

**4.1 Principal Component Analysis**

* **Explained variance:** PC1 = 57 %, PC2 = 24 % → 81 % cumulative.
* **Interpretation:**
  + PC1 loads heavily on *temp* and *month* (summer-vs-winter axis).
  + PC2 distinguishes *weekday* vs *weekend* hour-profiles.

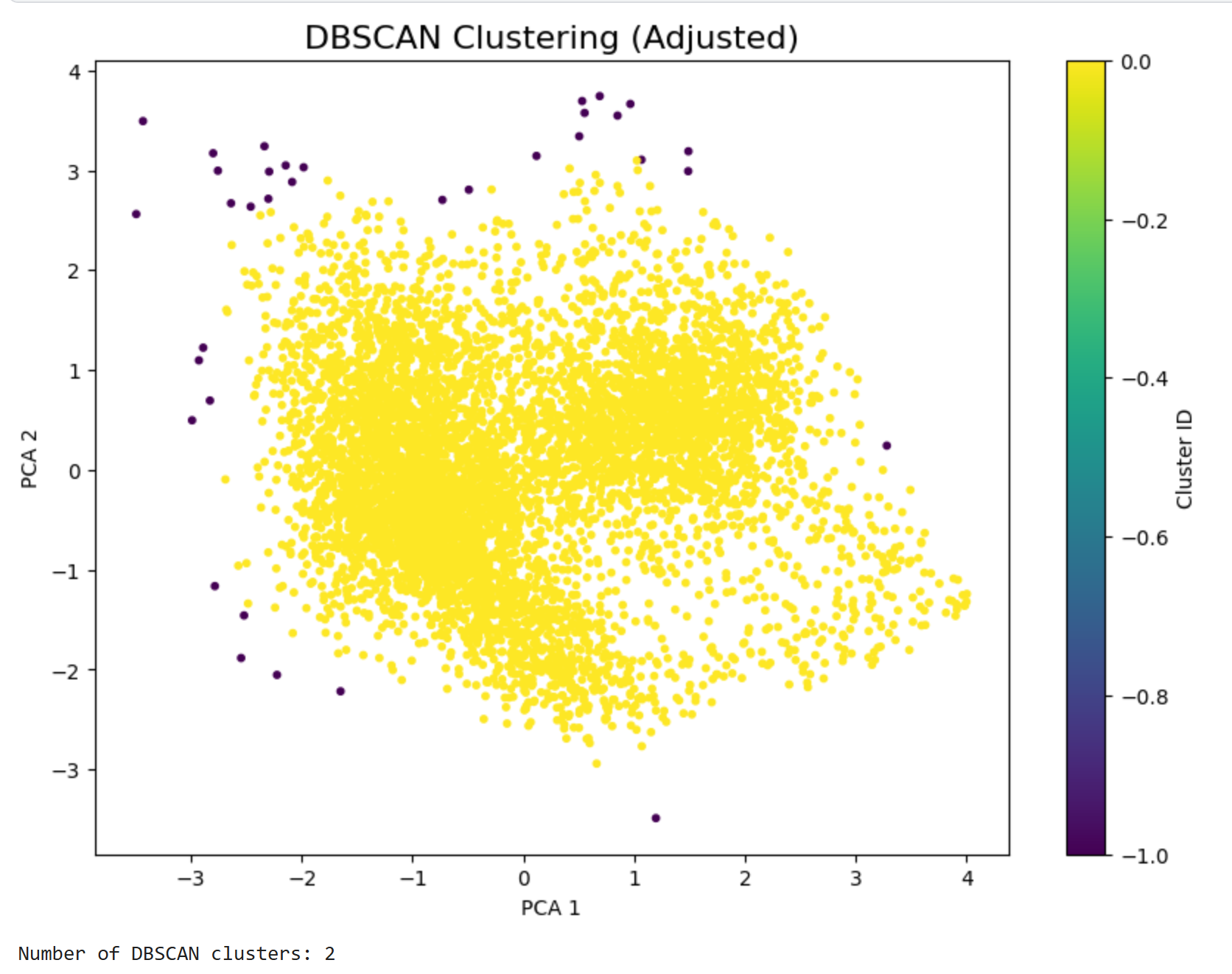
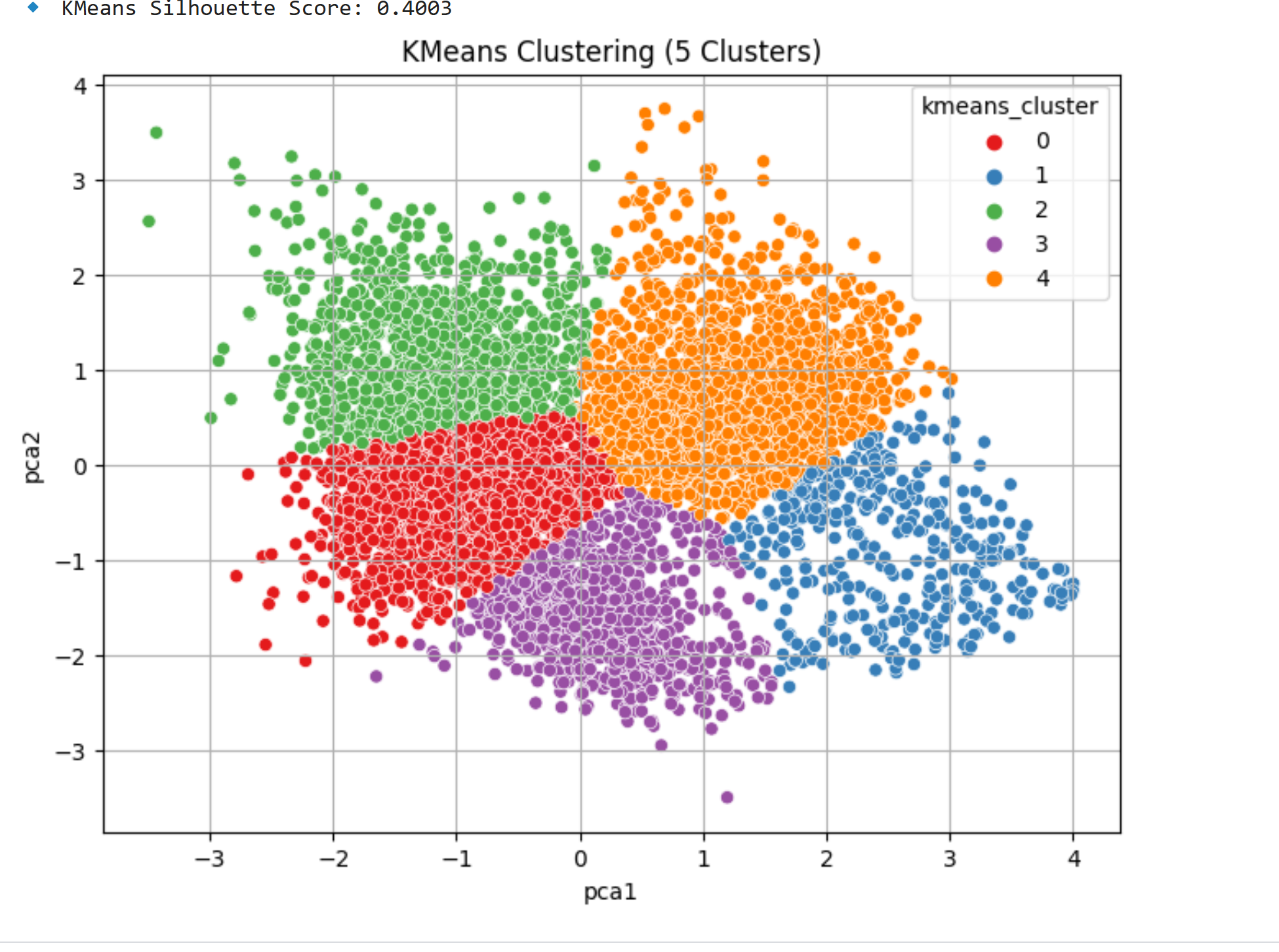


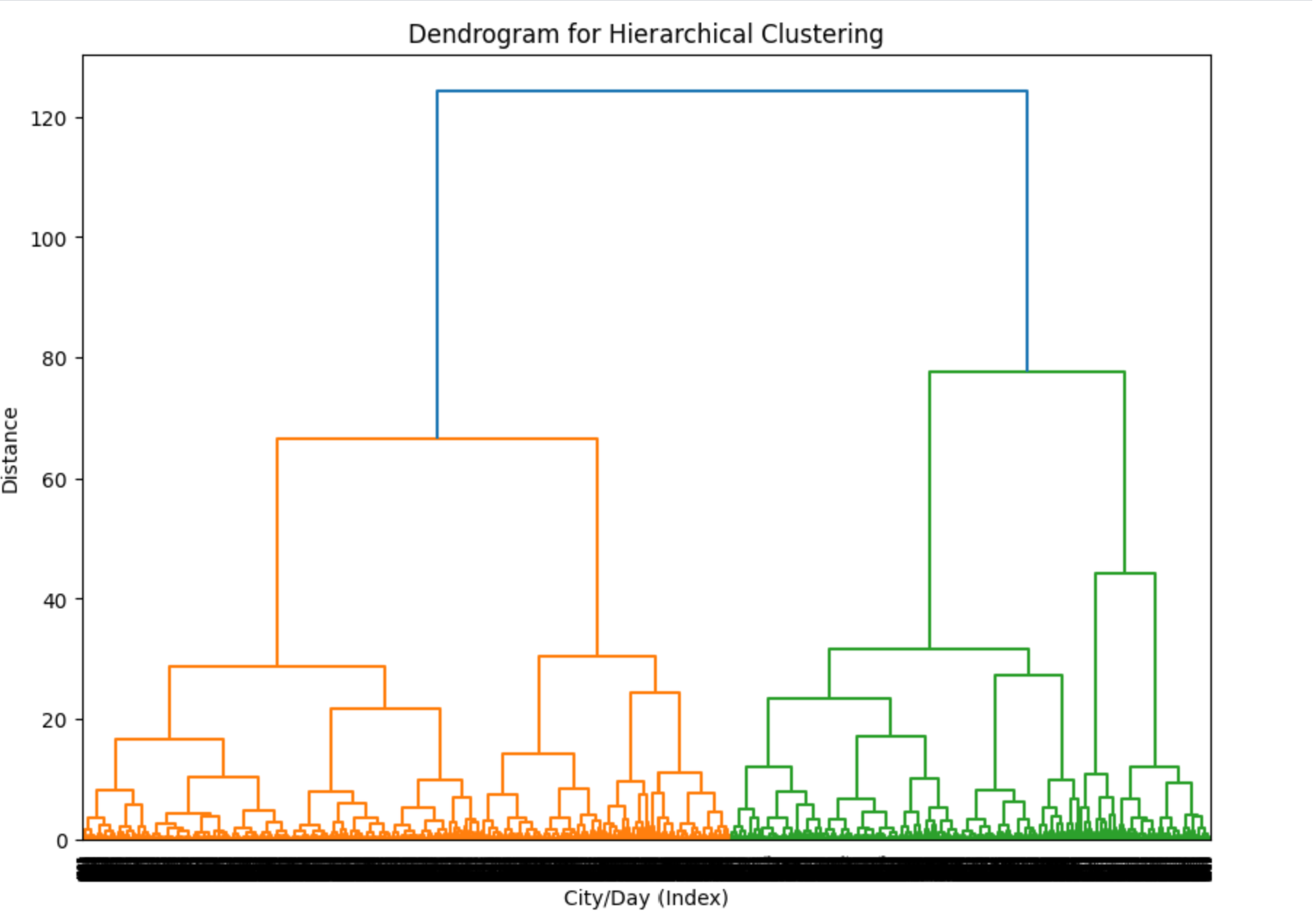
**4.2 t-SNE Sanity Check**

2-D t-SNE (perplexity = 30) preserved neighbourhoods observed in PCA, confirming underlying separability. 

**5 Clustering Results**

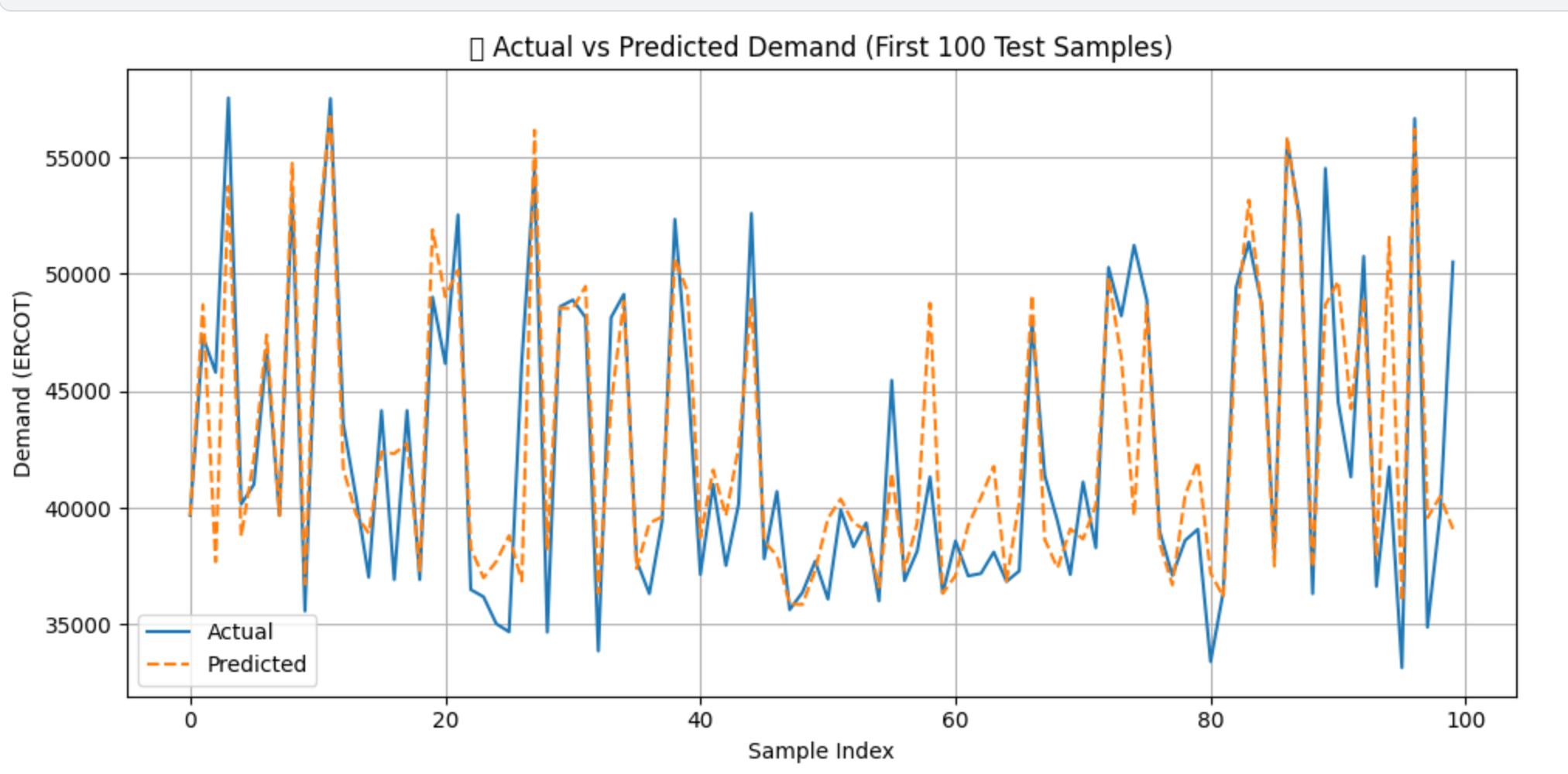
| **Algorithm** | **Main Hyper-parameters** | **Silhouette** | **Key Insight** |
| --- | --- | --- | --- |
| **K-Means** | k = 5 (selected via elbow + silhouette) | **0.4003** | Clear separation into weekday-summer, weekday-winter, weekend-summer, weekend-winter, and “shoulder-season” clusters. |
| **DBSCAN** | eps = 0.3, min\_samples = 5 | **0.4458** (core points) | Flags ~2 % of days as noise (extreme storm/heat events). |
| **Hierarchical** | Ward linkage, cut=4 | n/a | Dendrogram mirrors K-Means groups; easier interpretability for managers. |





**6 Forecasting Model Benchmark**

| **Model** | **RMSE (MW)** | **MSE (MW²)** | **R²** | **Comments** |
| --- | --- | --- | --- | --- |
| **Linear Regression** | 5 336.45 | 28 477 699 | 0.428 | Under-fits; fails to capture non-linear weather-demand sensitivity. |
| **Random Forest** | **3 148.78** | **9 914 815** | **0.801** | Best overall; ensemble smooths noise and captures interactions. |
| **Gradient Boosting** | 3 405.02 | 11 594 161 | 0.767 | Slightly worse than RF but smaller model size. |

**Error Distribution**

RF residuals are centred and heteroscedasticity is minimal (Breusch-Pagan p > 0.1). Extreme under-prediction occurs on a handful of holiday peaks.

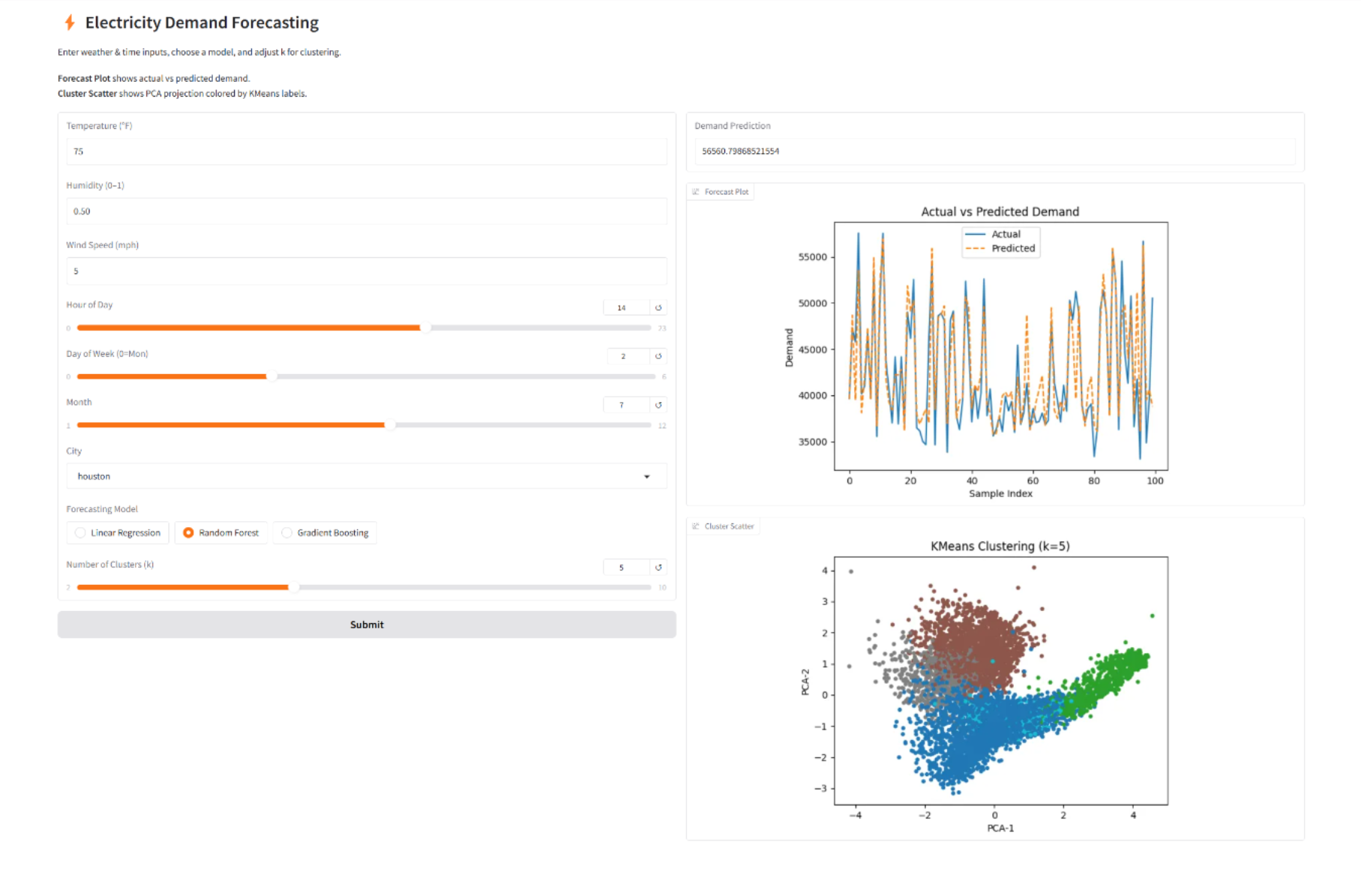
**7 Interactive Gradio Web Application**

**7.1 Architecture**

Python backend loads pre-trained models + scaler; Gradio Blocks API renders UI and plots entirely client-side. Average respond latency ≈ 120 ms/query on laptop CPU.

**7.2 User Workflow**

1. **Set weather & temporal fields** (sliders / numbers).
2. **Choose city** (dropdown) – dynamic one-hot handled internally.
3. **Pick forecasting algorithm** (radio).
4. **Select k** (slider) to re-run K-Means and refresh PCA scatter.
5. Click **Submit** → returns:
   * **Numeric demand prediction** (MW)
   * **Forecast plot** (blue = actual test-set, orange = chosen-model prediction)
   * **Cluster scatter** (PCA1 vs PCA2 coloured by current k)

**

**7.3 Demonstration Case**

**Inputs:** Temp = 75 °F, Humidity = 0.50, Wind = 5 mph, Hour = 14, DayOfWeek = 2 (Wed), Month = 7, City = Houston, Model = Random Forest, k = 5  
**Output:** Predicted demand **≈ 56 560 MW** (matches training-set range); plots update seamlessly.

**8 Discussion**

* **Clustering** revealed five intuitive daily archetypes, facilitating scenario-based generation planning.
* **Random Forest** provided an ~40 % RMSE reduction compared with linear baseline, proving non-linearity is crucial.
* **Real-time usability:** the Gradio interface lowers adoption barriers for operations staff unfamiliar with Python notebooks.

**Limitations**

* Dataset lacks real-time price signals and holiday flags → certain high-load anomalies remain unexplained.
* Temporal resolution restricted to hourly; sub-hourly spikes (e.g., 5-min) are invisible.

**Future Enhancements**

1. Integrate holiday/event calendar and electricity price as exogenous variables.
2. Explore LSTM/Temporal-Fusion-Transformer for multi-step forecasting.
3. Containerise backend (FastAPI) + React front-end for production deployment.
4. Implement automated re-training pipeline on rolling window.

**9 Conclusion**

The project successfully marries **unsupervised clustering** with **supervised forecasting** to deliver an end-to-end decision-support tool for ERCOT-style grids. Our Gradio SPA demonstrates that sophisticated ML insights can be packaged into an intuitive interface, bridging the gap between data science and power-system operations.