**Clustering Analysis of Utility Energy Consumption by ZIP Code**

Student: Jimmy Ma  
Course: CS777  
Date: February 24th 2025

1. Introduction

Utility energy consumption serves as a vital indicator of regional activity and efficiency. In densely populated states like New York, understanding variations in energy use can inform infrastructure planning, policy decisions, and targeted energy efficiency initiatives. This project employs clustering analysis to investigate whether distinct groups of ZIP codes in New York can be identified based solely on their monthly energy usage data. Using the publicly available Utility Energy Registry – Monthly ZIP Code Energy Usage dataset from New York’s Open Data Portal, the study seeks to uncover natural groupings and explore potential underlying factors—such as urbanization, economic activity, and seasonal influences—that differentiate these clusters.

2. Research Questions and Objectives

Primary Research Question:

* Can clustering analysis reveal distinct groups of ZIP codes based on their energy consumption patterns, and what underlying factors differentiate these groups?

Secondary Objectives:

* Determine whether ZIP codes can be segmented into groups representing low, moderate, high, and very-high energy consumption.
* Analyze the distribution of ZIP codes across these clusters to identify potential hotspots of high energy demand.
* Compare results from two clustering approaches—K-Means and Gaussian Mixture Model (GMM)—to evaluate their robustness and interpretability.
* Assess the potential drivers of consumption differences, including geographical factors and seasonal trends.

3. Data Description and Preprocessing

3.1 Data Source and Schema

The dataset used in this study is the Utility Energy Registry – Monthly ZIP Code Energy Usage dataset, available in CSV format from [New York’s Open Data Portal](https://data.ny.gov/Energy-Environment/Utility-Energy-Registry-Monthly-ZIP-Code-Energy-Us/tzb9-c2c6). The dataset includes:

* ZIP Code Information: Unique identifiers for each ZIP code.
* Energy Consumption (value): Monthly energy usage measurements.
* Temporal Data: Fields such as year and month.
* Additional Metadata: Fields like data\_class, data\_field, and others to provide contextual information.

A sample schema produced using PySpark is as follows:

root

|-- year: integer (nullable = true)

|-- data\_class: string (nullable = true)

|-- data\_field\_display\_name: string (nullable = true)

|-- data\_field: string (nullable = true)

|-- zip\_city: string (nullable = true)

|-- month: integer (nullable = true)

|-- zip\_code: integer (nullable = true)

|-- state\_2: string (nullable = true)

|-- uer\_id: integer (nullable = true)

|-- data\_stream: string (nullable = true)

|-- utility\_display\_name: string (nullable = true)

|-- value: double (nullable = true)

|-- number\_of\_accounts: integer (nullable = true)

|-- Georeference: string (nullable = true)

3.2 Preprocessing Steps

* Cleaning:  
  Rows with missing key fields (e.g., zip\_code and value) were removed. The energy usage field was explicitly cast to a double for numerical operations.
* Feature Engineering:  
  A single feature was constructed from the value field using PySpark’s Vector Assembler. No additional features were engineered at this stage, given the focus on energy consumption patterns.
* Data Splitting:  
  The cleaned dataset was split into training (70%) and testing (30%) subsets to ensure that the clustering model’s performance could be evaluated on unseen data.

4. Methodology

4.1 Clustering Algorithms

Two unsupervised learning techniques were used to partition the ZIP codes into distinct clusters:

K-Means Clustering

* Optimal k Determination:  
  The elbow method was applied by computing the within-cluster sum of squared errors (WSSSE) for k values ranging from 2 to 10. A custom function “find\_optimal\_k” measured the perpendicular distance from each point on the cost curve to the line connecting the first and last points, with the maximum distance indicating the optimal number of clusters. This process yielded an optimal k of 4.

WSSSE:

k = 2 -> Cost (WSSSE): 1.8075191589621676e+16

k = 3 -> Cost (WSSSE): 8515827506267837.0

k = 4 -> Cost (WSSSE): 5030313973413526.0

k = 5 -> Cost (WSSSE): 3393059452511035.0

k = 6 -> Cost (WSSSE): 2375133685017009.0

k = 7 -> Cost (WSSSE): 1819259560531832.2

k = 8 -> Cost (WSSSE): 1455066091651143.0

k = 9 -> Cost (WSSSE): 1192624712760309.8

k = 10 -> Cost (WSSSE): 947631217145830.4

A graph with a line

AI-generated content may be incorrect.

* Model Evaluation:  
  The K-Means model achieved a silhouette score of 0.9661 on the testing set, demonstrating excellent separation between clusters.

Gaussian Mixture Model (GMM)

* Probabilistic Clustering:  
  GMM represents each cluster as a Gaussian distribution and uses Expectation-Maximization (EM) to estimate parameters. With k=4 clusters (to allow direct comparison with K-Means), GMM produced a silhouette score of 0.9265.

4.2 Cluster Labeling and Visualization

After training both models, the cluster centers were sorted by their energy consumption values. Based on the sorted order, clusters were relabeled as:

* Low-Consumption Cluster
* Moderate-Consumption Cluster
* High-Consumption Cluster
* Very-High-Consumption Cluster

Scatter plots were generated to visualize the clustering. Each plot used a jitter value on the y-axis for visualization purposes, and vertical lines were drawn at the cluster center values. Heatmaps of binned energy usage further supported the interpretability of the clusters.

4.3 CSV Outputs

The final predictions from each model were saved to CSV files:

* kmeans\_energy\_cluster\_output.csv
* gmm\_energy\_cluster\_output.csv

Each CSV includes:

* zip\_code: The ZIP code identifier.
* value: The measured energy consumption.
* prediction: The numeric cluster assignment, later mapped to a descriptive label.

5. Results and Detailed Analysis

5.1 K-Means Clustering Results

Cluster Centers (Raw Values)

* Cluster 0: 10,940.83
* Cluster 1: 3,554,447.40
* Cluster 2: 651,134.94
* Cluster 3: 1,731,198.32

After sorting by increasing energy usage, the clusters were assigned as:

* Low-Consumption Cluster: 10,940.83
* Moderate-Consumption Cluster: 651,134.94
* High-Consumption Cluster: 1,731,198.32
* Very-High-Consumption Cluster: 3,554,447.40

CSV Output Insights (K-Means)

The K-Means CSV output reveals that:

* Majority Distribution: Most ZIP codes are grouped into the Low-Consumption and Moderate-Consumption clusters. This suggests that while the bulk of New York’s regions have moderate energy needs, a few areas exhibit substantially higher energy usage.
* Target Areas: The relatively small number of ZIP codes falling into the High-Consumption and Very-High-Consumption clusters likely represent densely populated urban or industrial zones. These regions may be prioritized for energy efficiency programs or targeted infrastructure investments.

Visualizations

* Scatter Plot: The scatter plot displays distinct, non-overlapping clusters with clear separation along the energy usage axis. Each cluster’s vertical marker (representing the center) helps validate the cluster boundaries.

A screen shot of a graph

AI-generated content may be incorrect.

* Heatmap: The heatmap confirms that energy usage frequencies vary significantly by cluster, further validating the effectiveness of the clustering.

A screenshot of a computer

AI-generated content may be incorrect.

5.2 GMM Clustering Results

Cluster Centers (Raw Values)

* Cluster 0: 256,325.05
* Cluster 1: 2,840.66
* Cluster 2: 2,488,824.70
* Cluster 3: 1,001,461.68

After sorting:

* Low-Consumption Cluster: 2,840.66
* Moderate-Consumption Cluster: 256,325.05
* High-Consumption Cluster: 1,001,461.68
* Very-High-Consumption Cluster: 2,488,824.70

CSV Output Insights (GMM)

The GMM CSV output, while showing a similar overall segmentation, presents slight differences:

* Distribution Consistency: As with K-Means, most ZIP codes fall into the Low- and Moderate-Consumption groups.
* Model Sensitivity: The small discrepancies between the K-Means and GMM cluster centers (and subsequent ZIP code assignments) highlight the sensitivity of clustering outcomes to the underlying model assumptions. GMM’s probabilistic approach captures uncertainty and overlapping cluster boundaries, which may be beneficial for certain downstream analyses.

Visualizations

* Scatter Plot and Heatmap: Like the K-Means results, the scatter plot for GMM displays distinct clusters with well-defined centers. The heatmap reinforces the energy usage distribution across clusters, corroborating the CSV output findings.

A diagram of a red and green chart

AI-generated content may be incorrect.

A grid of numbers and letters

AI-generated content may be incorrect.

5.3 Comparative Analysis: K-Means vs. GMM

Methodological Differences

* K-Means Clustering:
  + Hard Assignments: K-Means assigns each data point definitively to the nearest cluster center based on Euclidean distance. This results in crisp boundaries between clusters.
  + Spherical Clusters: The algorithm assumes clusters are spherical and of roughly similar size, which is suitable when the data’s variance is uniform.
  + Optimization Objective: It minimizes the within-cluster sum of squared errors (WSSSE), yielding very high silhouette scores when clusters are well-separated. In this project, K-Means achieved a silhouette score of 0.9661, indicating highly distinct clusters.
* Gaussian Mixture Model (GMM):
  + Soft Assignments: GMM assigns each data point a probability of belonging to each cluster. This probabilistic approach can better capture overlapping or ambiguous data points.
  + Elliptical Clusters: GMM does not restrict clusters to be spherical; it can model elliptical clusters and account for varying shapes and variances.
  + Model Complexity: By using a mixture of Gaussians, GMM incorporates the uncertainty in cluster assignments. This can result in slightly lower silhouette scores compared to K-Means, as seen in this project with a score of 0.9265.

Cluster Center Characteristics

* K-Means Cluster Centers (Raw Values):
  + ~10,941, ~651,135, ~1,731,198, and ~3,554,447
  + Interpretation: These centers clearly delineate the four consumption tiers, with low energy usage in one cluster and a dramatic jump to very high usage in another. The crisp assignment reinforces distinct groupings.
* GMM Cluster Centers (Raw Values):
  + ~2,840, ~256,325, ~1,001,462, and ~2,488,825
  + Interpretation: The centers from GMM tend to be shifted slightly lower or higher in some cases relative to K-Means. This difference arises because GMM’s probabilistic model weighs data points differently, especially in regions where clusters overlap. For instance, the Low-Consumption Cluster in GMM is centered around ~2,840 compared to ~10,941 in K-Means—highlighting that the hard boundary of K-Means may overestimate the minimum consumption level.

Distribution of Cluster Assignments

* Assignment Consistency:
  + Both methods segment most ZIP codes into the Low-Consumption and Moderate-Consumption clusters, with fewer ZIP codes in the High and Very-High-Consumption clusters.
  + K-Means tends to produce sharper boundaries, meaning that each ZIP code is assigned unequivocally to one cluster based solely on its energy usage value.
  + GMM’s probabilistic assignments allow some ZIP codes to exhibit partial membership in multiple clusters. This nuance can be useful when considering borderline cases or when further integrating external factors (e.g., demographic or economic data) to better understand overlapping usage patterns.

Silhouette Score Comparison

* K-Means: With a silhouette score of 0.9661, the clusters are highly compact and well-separated. This suggests that the clear-cut assignments of K-Means match the natural groupings in the data.
* GMM: Achieving a score of 0.9265, GMM also forms good clusters but with a slightly lower score—reflecting its capacity to handle overlapping clusters. The marginally lower score indicates that, while the clusters are still distinct, there is more ambiguity in the boundaries.

In summary, both clustering models reveal four meaningful consumption segments. The slight differences in cluster centers and silhouette scores highlight these methodological contrasts and offer complementary insights into New York’s energy consumption patterns.

6. Discussion

6.1 Key Observations

* Dominant Clusters:  
  Most ZIP codes are in the Low-Consumption and Moderate-Consumption clusters, indicating that extreme energy usage is relatively uncommon. This pattern may reflect widespread residential use with moderate commercial activity.
* Outlier Regions:  
  The High-Consumption and Very-High-Consumption clusters, though representing fewer ZIP codes, likely correspond to urban centers or industrial zones. These areas warrant further investigation to determine if higher energy usage stems from economic activity, population density, or other factors.
* Model Comparison:  
  While both K-Means and GMM successfully segment the data, the marginally higher silhouette score for K-Means suggests its clusters are more distinct. GMM’s probabilistic clustering, however, provides a nuanced perspective on data overlap and uncertainty, which could be advantageous in scenarios where ZIP codes border different consumption patterns.

6.2 Implications for Policy and Utility Management

* Targeted Interventions:  
  Identifying ZIP codes in the High-Consumption and Very-High-Consumption clusters can help utility providers focus on energy conservation and load management programs. Similarly, policymakers can prioritize infrastructure investments in these areas.
* Further Data Integration:  
  The CSV outputs serve as a foundation for integrating additional data sources. Merging with demographic, economic, or weather-related data could enhance the understanding of the drivers behind energy consumption and lead to more effective interventions.

7. Conclusion

The project demonstrates that clustering analysis is an effective tool for segmenting ZIP codes in New York by their energy consumption patterns. Both K-Means and GMM identified four distinct clusters—corresponding to low, moderate, high, and very-high energy usage—with high silhouette scores validating the robustness of these groupings. The analysis of CSV output further confirms that while most areas have moderate energy demands, a small subset of regions exhibit significantly higher consumption. These findings suggest that underlying factors such as urban density, industrial activity, and seasonal variations are likely driving the observed differences.

In summary:

* Clustering reveals meaningful segments: The energy consumption data naturally clusters into four distinct groups.
* Actionable insights: The segmentation can inform targeted energy efficiency programs and infrastructure investments.
* Future directions: Integrating additional features and extending the analysis temporally will further refine these insights.