Optimizing Data Center Location Selection
A Strategic Approach Using Clustering, Gravity Modeling & NSGA-II

Submitted by:
Mahima Advilkar
Abhinav Krishnamoorthy
Prasad Pilankar
Navin Kishore Ravi Kumar

Program: MS Business Analytics, CSU East Bay
Course: Capstone Project, Spring 2025

Capstone Project GitHub Link:

Table of Contents

1. Introduction / Business Problem

In today's digitally driven economy, companies rely heavily on data centers to support operations, store information, and ensure seamless service delivery. However, inefficient location strategies and suboptimal operational planning often result in increased operational costs, higher latency, and compliance issues. This capstone project addresses the critical business challenge: “How can companies strategically choose locations and optimize operational efficiency for their data centers to balance costs, performance, compliance, and scalability?”

The remainder of this report is structured as follows: Section 2 reviews related literature on data center optimization. Section 3 outlines the data sources, attributes, and preprocessing steps. Section 4 describes the analytical models employed, including clustering, gravity modeling [11], and NSGA-II [10]. Section 5 discusses results and interpretations. Section 6 presents conclusions and business impact. Section 7 details team contributions. Section 8 contains references, and Section 9 includes appendix code, tables, and figures.

2. Related Work

This project builds on scholarly and industry research surrounding data center location optimization and operational modeling. Studies such as Buyya et al. [2] present cloud resource scheduling strategies, while Deb et al. [1] introduced the NSGA-II framework for multi-objective optimization. Additionally, Ahmed et al. [3] proposed energy-aware site selection for green data centers. Liu et al. [10] conducted a multi-objective optimization study for data center location and resource allocation in China using clustering and genetic algorithms, providing useful methodological parallels.

Industry references such as AWS [4] and IBM [5] highlight best practices for scalable data infrastructure. Gravity models have been used in logistics and urban planning [9] and recently extended to infrastructure planning. Our project uniquely integrates clustering, gravity modeling [11], and NSGA-II [10] to analyze real-world data and deliver location-specific, optimization-focused insights for data center planning.

3. Data Description

The dataset utilized in this analysis was primarily obtained through web scraping techniques applied to trusted industry sources such as DataCenterMap [8], government open-data APIs (e.g., IEA) [6], and official data center provider websites such as AWS [4] and IBM [5]. Data was systematically scraped, aggregated, and integrated to ensure comprehensive geographic coverage, including more than 50 existing data centers across the United States and 13 major Indian cities evaluated for future expansion.

- **Data Sources:** DataCenterMap [8], governmental open data platforms, provider-specific websites, Wikipedia.
- **Geographic Scope:** United States (50+ data centers), India (13 cities).
- **Key Attributes:** ENERGY (MW), AREA (sq ft), IT POWER, CABINETS availability, PUE, IXPs, Facility Age, Infrastructure quality, Land costs, Climate considerations.
- **Data Processing:** Duplicate removal, feature normalization, missing value imputation based on industry standards.
- **Derived Metrics:** Infrastructure composite scores, service availability indices, standardized cost indicators.

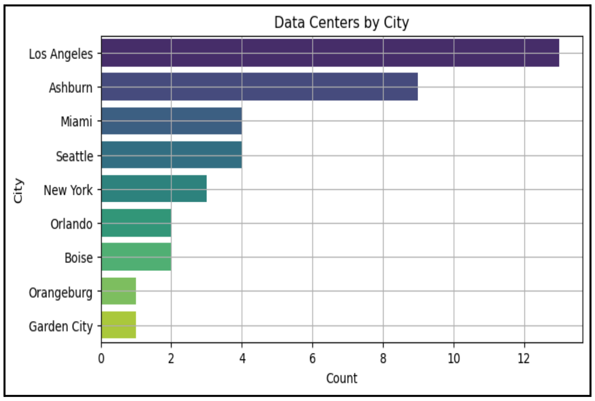


Figure A1: Distribution of data centers across major U.S. cities, showing Ashburn and Los Angeles as leading hubs

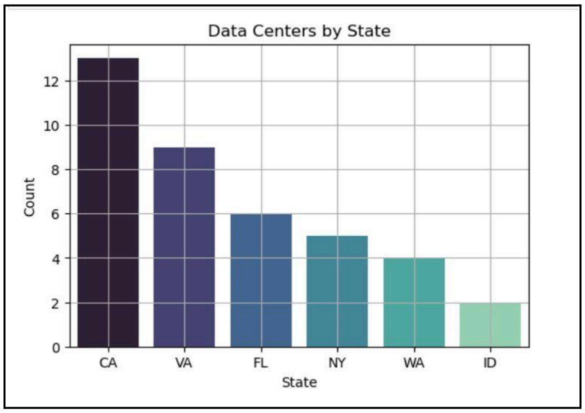
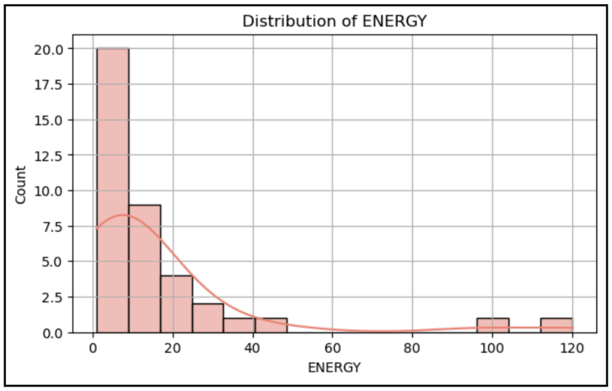


Figure A2: California leads with 13 data centers, followed by Virginia with 9, Florida with 6, New York with 5, Washington with 4, and Idaho with 2—indicating key infrastructure clusters on both coasts.

Figure A3: Distribution of ENERGY Most data centers (20+) operate under 20 MW, with a long right-skewed tail extending up to 120 MW, indicating a few large-scale outliers



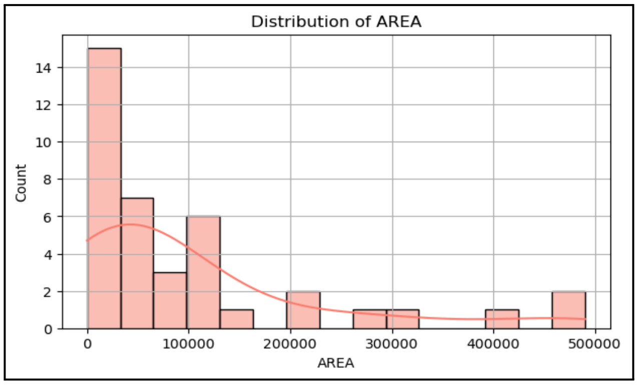


Figure A4: Nearly 15 centers are under 100,000 sq ft, but the area distribution is highly skewed, with some exceeding 450,000 sq ft, revealing significant facility size variance.

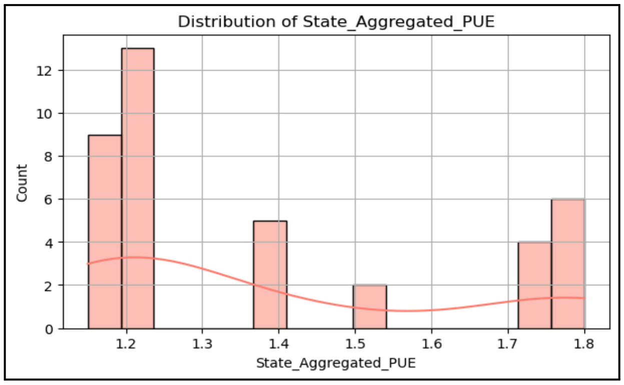


Figure A5: The majority of centers report PUE between 1.18 and 1.30, with peaks around 1.20, suggesting generally efficient energy usage; a few inefficient centers reach up to 1.76.

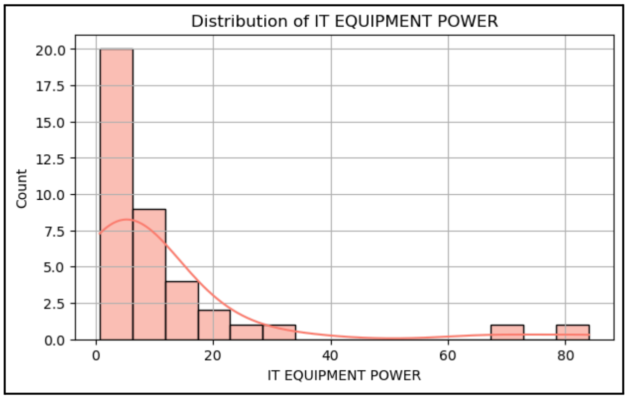


Figure A6: (IT POWER):Over 60% of centers use less than 15 MW; some reach 85 MW capacity.

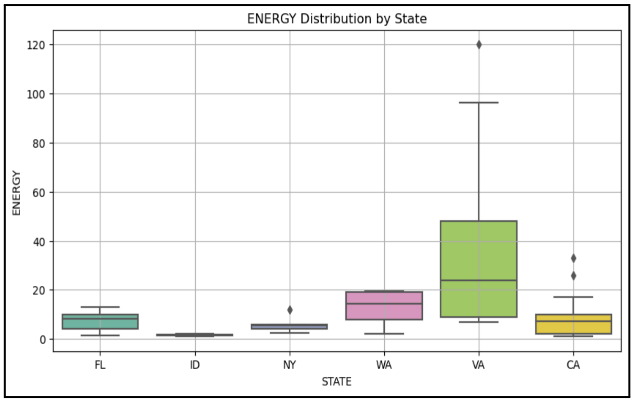
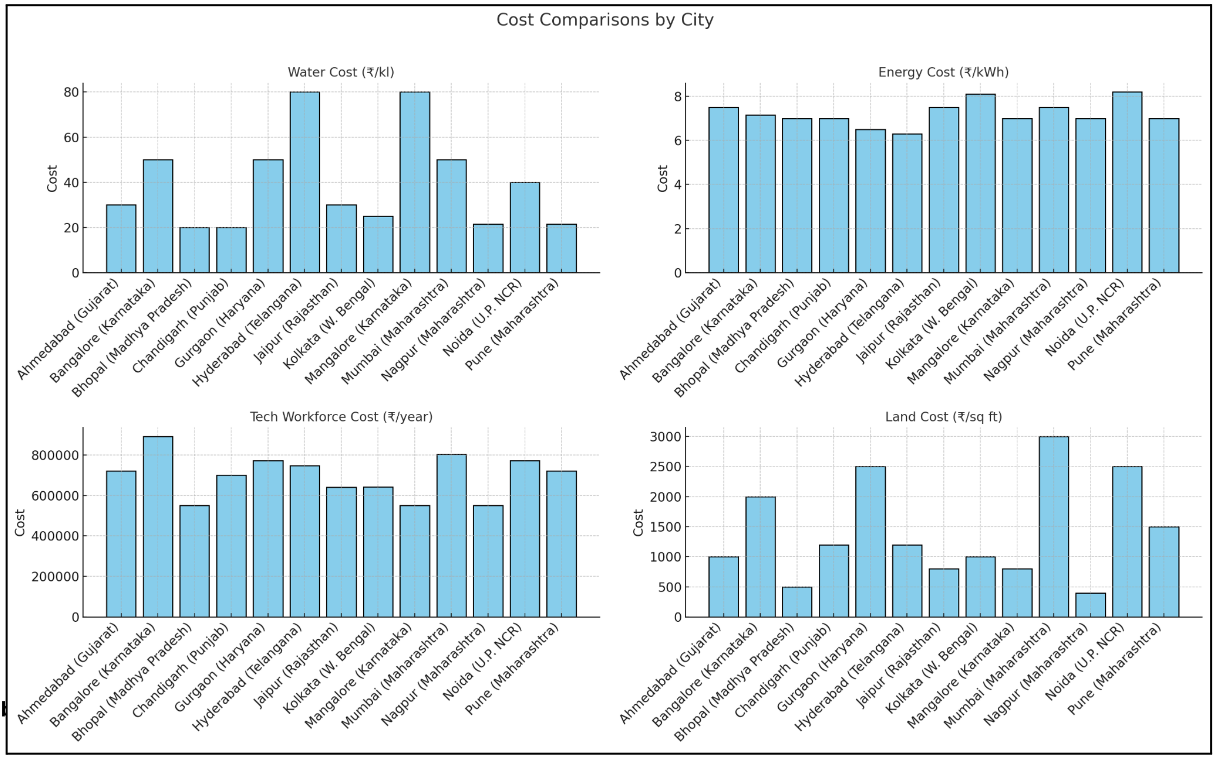


Figure A7: Virginia shows a wide energy range; Idaho and New York remain tightly clustered under 5 MW.



Tal

Figure A8:

- **Mumbai reports the highest land cost (~₹3,000/sq ft), while Bhopal and Nagpur are lowest.**
- **Bangalore leads tech salaries (~₹900K/year); Mangalore and Bhopal offer lowest wages, highlighting labor cost gaps.**
- **Energy costs range from ₹6.5–₹8.2/kWh, with Nagpur and Kolkata among the most expensive cities.**
- **Nagpur and Jaipur have the highest water tariffs (~₹80/kl); Chandigarh, Pune, Noida are lowest.**

Key Variables: Water Cost (₹ per kiloliter), Energy Cost (₹ per kWh) Tech Workforce Cost (Average Annual Salary in ₹), Land Cost (₹ per square foot)

Table A1: Descriptive Stats for Indian Data Center Market

	Water	Energy	Workforce	Renewable	LandCost	LandAvail	Network	Climate
Count	13.00	13.00	13	13.00	13.00	13.00	13.00	13.00
Mean	41.27	7.23	696538	44.84	1421.15	6.85	7.00	7.15
STD	22.55	0.52	105783	19.63	823.09	1.68	1.63	1.07
Min	20.00	6.30	550000	13.50	475.00	3.00	5.00	5.00
25%	21.50	7.00	640000	32.00	800.00	6.00	6.00	7.00
50%	30.00	7.00	720000	45.00	1200.00	7.00	6.00	7.00
75%	50.00	7.50	771000	59.50	2000.00	8.00	8.00	8.00
Max	80.00	8.20	891000	74.60	3000.00	9.00	10.00	9.00

Note: Data Collection for Indian Market is solely from desk/primary research and the published articles and sources on the internet. The above link presents the in depth details on data collection for this gravity model performance. Data Source for Indian Market and Raw Data Source File

Key attributes collected include:

- **ENERGY (MW):** Power consumption capacity.
- **AREA (sq ft):** Facility size.
- **IT EQUIPMENT POWER:** Computing power utilization.
- **Infrastructure Features:** Availability indicators such as cabinets, cages, suites, remote hands.
- **Power Usage Effectiveness (PUE):** Operational efficiency metric.
- **Internet Exchange Points (IXPs):** Connectivity and network reliability metrics.
- **Facility Age:** Years operational, indicating technological relevance.
- **Cost Metrics:** Land acquisition, construction, and operational expenditures.
- **Climate and Sustainability Indices:** Environmental impact and sustainability benchmarks.

The data underwent a rigorous **ETL (Extract, Transform, Load)** process, including removal of duplicates, normalization of numerical features, and filling of missing values using standardized industry benchmarks. Additionally, derived attributes like composite infrastructure scores and service availability indices were computed to facilitate sophisticated analytical modeling such as clustering, gravity modeling, and multi-objective optimization.

4. Analytical Models and Methodology

4.1 Clustering and PCA

Objective:

To segment existing U.S. data centers into **distinct operational clusters** based on performance, infrastructure, and efficiency metrics. This classification enables **benchmarking**, **predictive planning**, and **targeted investment** decisions

Methodology:

- **Feature Selection and Scaling:** Relevant numerical and categorical operational features were selected (ENERGY, AREA, IT Equipment Power, Infrastructure Availability, PUE, IXPs). All features were standardized using StandardScaler to prevent feature dominance due to scale differences.

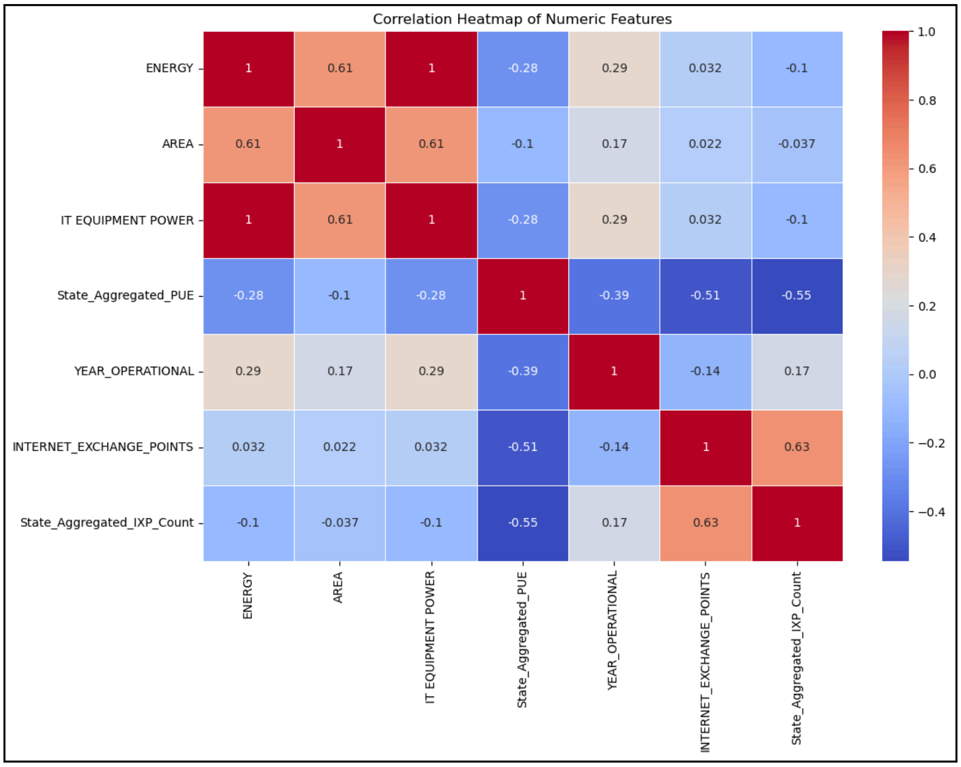


Figure A9: Correlation Heatmap of Numeric Features

The heatmap shows that ENERGY, AREA, and IT POWER are strongly correlated, confirming that larger data centers consume more power. PUE is negatively correlated with connectivity features like IXPs, suggesting that well-connected centers are generally more energy efficient. These insights validated our feature selection and supported the use of PCA to handle multicollinearity.

- **Clustering (KMeans):** Applied the KMeans algorithm with an optimal cluster number (k=3) determined through iterative analysis and cluster validity measures.

`

- **Dimensionality Reduction (PCA):** Principal Component Analysis (PCA) was employed for visual representation of the clusters, significantly reducing data dimensionality while preserving approximately 70% variance with two principal components.

Cluster Profiles:1

- **Cluster 0 (Medium-sized Efficient Centers):**

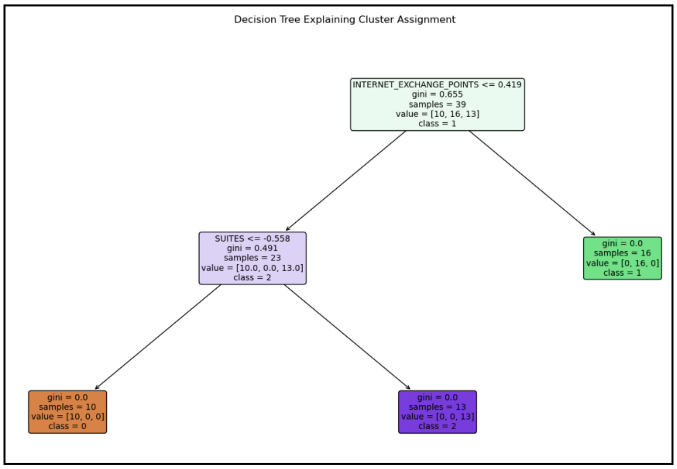
    a. Avg. Energy Usage: 8.75 MW,
    b. Avg. Facility Area: 82,346 sq ft,
    c. Avg. PUE: 1.25,
    d. Comprehensive infrastructure support; mature,well-connected centers.Well-balanced in efficiency and service features. Ideal for replication in enterprise deployments

- **Cluster 1 (Large-scale Basic Centers):**

    a. Avg. Energy Usage: 36.9 MW
    b. Avg. Facility Area: 178,514 sq ft
    c. Avg. PUE: 1.21 (high efficiency)
    d. Minimal infrastructure services; newer facilities ideal for large wholesale or hyperscale operations.

- **Cluster 2 (Small-scale Inefficient Centers):**

    a. Avg. Energy Usage: 7.23 MW
    b. Avg. Facility Area: 68,161 sq ft
    c. Avg. PUE: 1.67 (low efficiency)
    d. Older centers with extensive infrastructure services but significant efficiency improvement needed.



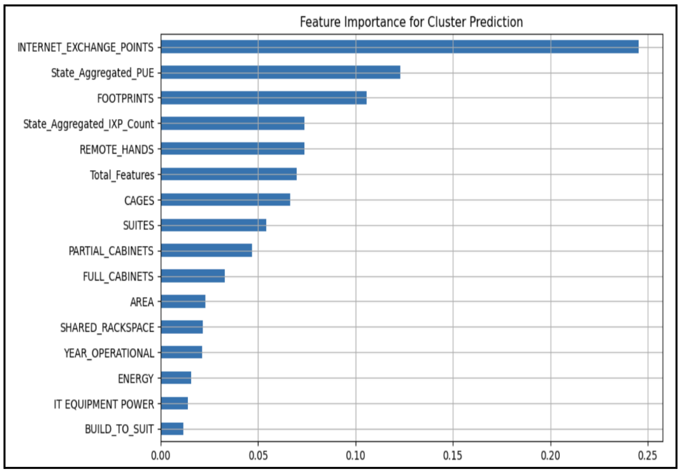


Figure A10: Decision Tree for Cluster Assignment
This helped visualize how decisions could be made hierarchically, making the segmentation transparent and interpretable for business stakeholders. For example, a data center with low IXPs and limited infrastructure would fall into Cluster 1 (large, low-service centers), guiding both expansion and optimization decisions

Figure A11: Feature Importance for Clustering
This ranking provides quantitative evidence for which features matter most in defining operational types—helping leaders prioritize the right factors (e.g., connectivity, energy efficiency) during new data center planning.

Purpose and Business Application:
This clustering framework converts unstructured facility data into actionable categories. It enables:

- **Benchmarking:** Map existing or proposed centers to a cluster to evaluate performance alignment.
- **Investment Strategy:** Identify centers for upgrade (Cluster 2) or replicate successful designs (Clusters 0 & 1).
- **Predictive Planning:** A Random Forest classifier was trained on these clusters (see Section 5.2) to assign new centers to a cluster based on specs—supporting early-stage strategy and risk mitigation.

Clustering was used to segment U.S. data centers into three meaningful operational groups based on performance and infrastructure features. This **classification** enables better benchmarking, smarter resource allocation, and predictive planning. It helps business teams decide **where to invest, which facilities to optimize, and how to design future centers with cost-efficiency and performance in mind.**

In essence, clustering transforms raw operational data into a strategic classification system, helping businesses make informed decisions around upgrades, expansion, and capacity planning.

4.2 Gravity Model

$$\boxed{\text{Score}_i = \sum_j (w_j \cdot \text{norm}(x_{ij}))}$$

Where:

- $Score_i$ = Final gravity score for city *i*
- w_j = Weight assigned to parameter *j*
- $norm(x_i j)$ = Normalized value of parameter *j* for city *i*

Objective:
Identify and rank potential Indian city locations based on attractiveness for future data center deployment, considering multiple strategic factors such as costs, connectivity, sustainability, and scalability.

Methodology:
A Gravity Model was employed, leveraging weighted scoring of key criteria influencing site selection using the data from [**Table A2]**. Each city's attractiveness score was calculated using the following dimensions:

- **Cost Factors:** Energy, land acquisition, water, and workforce costs.
- **Connectivity:** Network infrastructure and proximity to Internet Exchange Points (IXPs).
- **Renewable Energy and Sustainability:** Availability and usage percentage of renewable energy sources.
- **Land Availability:** Scalability potential based on available land.
- **Climate Resilience:** Assessment of environmental conditions affecting operational risk.

Weights Applied:

- Cost (Energy, Land, Workforce, Water): 40%
- Connectivity (Network, IXPs): 20%
- Renewable Energy and Sustainability: 20%
- Scalability and Expansion Potential (Land Availability, Climate): 20%

Actual Formula:

$$\boxed{\begin{gathered} Score_i = 0.05 * norm(Water_i) + 0.20 * norm(Energy_i) + 0.05 * norm(Workforce_i) + \\ 0.10 * norm(Renewable_i) + 0.15 * norm(LandCost_i) + 0.05 * norm(LandAvail_i) + \\ 0.20 * norm(Network_i) + 0.10 * norm(Climate_i) \end{gathered}}$$

Results:
Hyderabad emerged as the top-ranked city with the highest composite attractiveness score, driven primarily by optimal energy costs (6.3 ₹/kWh), robust network connectivity (connectivity index 8/10), high renewable energy potential (42.9%), and strong scalability prospects. Bangalore and Pune followed closely, also reflecting favorable conditions in cost efficiency, connectivity, and sustainable practices.

This gravity modeling approach provided strategic clarity, helping decision-makers prioritize potential locations systematically and objectively, ensuring alignment with long-term operational goals and sustainability commitments.

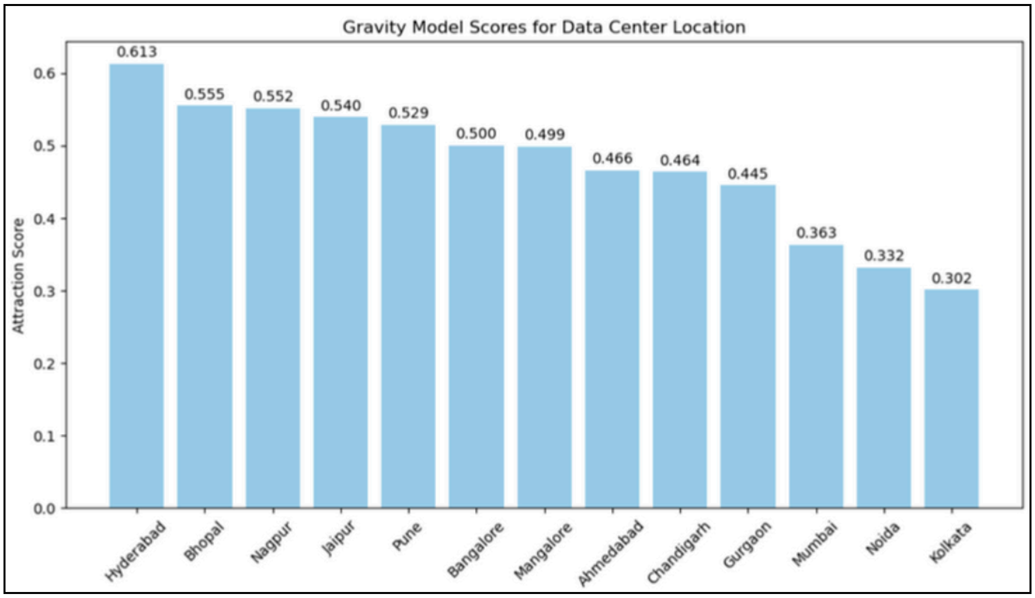


Figure A12: Gravity Model Scores for Data Center Location

Hyderabad ranks highest in attraction score (0.613), followed by Bhopal, Nagpur, and Jaipur. Kolkata, Noida, and Mumbai score the lowest, indicating relatively weaker suitability for data center expansion based on weighted infrastructure, cost, and efficiency indicators.

4.3 NSGA-II Multi-Objective Optimization

$$\boxed{A_i = \sum_{k=1}^{n} w_k \cdot f_{ik}}$$

Where:

$$\begin{aligned} A_i &= \text{Aggregated (weighted) score for solution } i \\ w_k &= \text{Weight assigned to objective } k \text{ (user-defined)} \\ f_{ik} &= \text{Normalized value of the } k^{\text{th}} \text{ objective in solution } i \\ n &= \text{Total number of objectives (in your case, } n = 4) \end{aligned}$$

Objective:

To optimize data center site selection by effectively managing and balancing multiple conflicting operational objectives, enabling decision-makers to achieve the most beneficial strategic outcomes.

Methodology:

The Non-dominated Sorting Genetic Algorithm II (NSGA-II) was employed for multi-objective optimization. This advanced optimization method simultaneously addresses conflicting criteria by generating a set of optimal solutions known as the Pareto front.

Optimization Objectives:

- **Minimize Power Usage Effectiveness (PUE):**
  Improve energy efficiency and reduce operational costs.

$$\boxed{f_1 = \sum_{i \in S} PUE_i}$$

- **Maximize Connectivity (IXP Count):**
  Ensure robust network performance and minimize latency.

$$\boxed{f_2 = -\sum_{i \in S} IXP_i}$$

- **Maximize Infrastructure and Service Availability Score:**
  Offer comprehensive infrastructure capabilities for diverse operational needs.

$$\boxed{f_3 = -\sum_{i \in S} ServiceScore_i}$$

- **Minimize Facility Age:**
  Ensure facilities are modern and capable of supporting advanced technology and future scalability.

$$\boxed{f_4 = \frac{1}{|S|} \sum_{i \in S} Age_i}$$

Constraints Applied:

- Minimum IT Power Capacity: 1 MW
- Minimum Facility Area: 10,000 sq ft

$$\boxed{Valid(i) = \begin{cases} 1 & \text{if } Power_i \geq 1 \land Area_i \geq 10000 \\ 0 & \text{otherwise} \end{cases}}$$

Final Formula:

$$\boxed{A_i = -w_1 \cdot \widehat{f_1(i)} + w_2 \cdot \widehat{f_2(i)} + w_3 \cdot \widehat{f_3(i)} - w_4 \cdot \widehat{f_4(i)}}$$

With weights: $w_1 = 0.4,\ w_2 = 0.3,\ w_3 = 0.2,\ w_4 = 0.1$

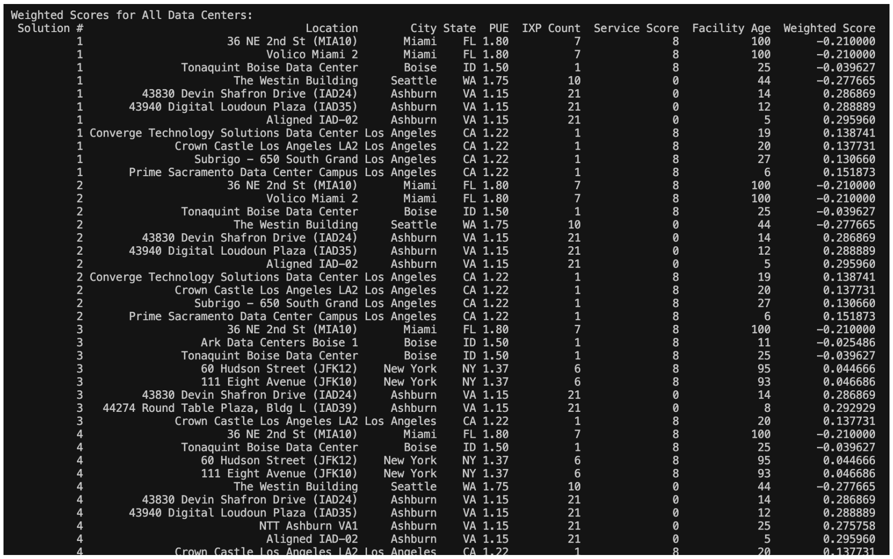
Where: $\widehat{f_k(i)} = \text{Normalized value of objective } f_k \text{ for solution } i$

$w_k = \text{User-defined weight for objective } k \quad (w_1 + w_2 + w_3 + w_4 = 1)$

The weights considered here are in the order of what makes data center efficiency tick and general objectives; however, it can be tweaked based on what is the most important between PUE, Internet Connectivity, Service Availability and Facility Age.

Results:

The NSGA-II model produced clear Pareto-optimal solutions, illustrating optimal trade-offs between efficiency, connectivity, infrastructure, and facility age.



```
Weighted Scores for All Data Centers:
 Solution #                                  Location        City State  PUE IXP Count Service Score Facility Age Weighted Score
          1                      36 NE 2nd St (MIA10)       Miami    FL 1.80         7             8          100      -0.210000
          1                            Volico Miami 2       Miami    FL 1.80         7             8          100      -0.210000
          1               Tonaquint Boise Data Center       Boise    ID 1.50         1             8           25      -0.039627
          1                       The Westin Building     Seattle    WA 1.75        10             0           44      -0.277665
          1         43830 Devin Shafron Drive (IAD24)     Ashburn    VA 1.15        21             0           14       0.286869
          1       43940 Digital Loudoun Plaza (IAD35)     Ashburn    VA 1.15        21             0           12       0.288889
          1                            Aligned IAD-02     Ashburn    VA 1.15        21             0            5       0.295960
          1 Converge Technology Solutions Data Center Los Angeles    CA 1.22         1             8           19       0.138741
          1              Crown Castle Los Angeles LA2 Los Angeles    CA 1.22         1             8           20       0.137731
          1                 Subrigo - 650 South Grand Los Angeles    CA 1.22         1             8           27       0.130660
          1       Prime Sacramento Data Center Campus Los Angeles    CA 1.22         1             8            6       0.151873
          2                      36 NE 2nd St (MIA10)       Miami    FL 1.80         7             8          100      -0.210000
          2                            Volico Miami 2       Miami    FL 1.80         7             8          100      -0.210000
          2               Tonaquint Boise Data Center       Boise    ID 1.50         1             8           25      -0.039627
          2                       The Westin Building     Seattle    WA 1.75        10             0           44      -0.277665
          2         43830 Devin Shafron Drive (IAD24)     Ashburn    VA 1.15        21             0           14       0.286869
          2       43940 Digital Loudoun Plaza (IAD35)     Ashburn    VA 1.15        21             0           12       0.288889
          2                            Aligned IAD-02     Ashburn    VA 1.15        21             0            5       0.295960
          2 Converge Technology Solutions Data Center Los Angeles    CA 1.22         1             8           19       0.138741
          2              Crown Castle Los Angeles LA2 Los Angeles    CA 1.22         1             8           20       0.137731
          2                 Subrigo - 650 South Grand Los Angeles    CA 1.22         1             8           27       0.130660
          2       Prime Sacramento Data Center Campus Los Angeles    CA 1.22         1             8            6       0.151873
          3                      36 NE 2nd St (MIA10)       Miami    FL 1.80         7             8          100      -0.210000
          3                  Ark Data Centers Boise 1       Boise    ID 1.50         1             8           11      -0.025486
          3               Tonaquint Boise Data Center       Boise    ID 1.50         1             8           25      -0.039627
          3                  60 Hudson Street (JFK12)    New York    NY 1.37         6             8           95       0.044666
          3                  111 Eight Avenue (JFK10)    New York    NY 1.37         6             8           93       0.046686
          3         43830 Devin Shafron Drive (IAD24)     Ashburn    VA 1.15        21             0           14       0.286869
          3   44274 Round Table Plaza, Bldg L (IAD39)     Ashburn    VA 1.15        21             0            8       0.292929
          3              Crown Castle Los Angeles LA2 Los Angeles    CA 1.22         1             8           20       0.137731
          4                      36 NE 2nd St (MIA10)       Miami    FL 1.80         7             8          100      -0.210000
          4               Tonaquint Boise Data Center       Boise    ID 1.50         1             8           25      -0.039627
          4                  60 Hudson Street (JFK12)    New York    NY 1.37         6             8           95       0.044666
          4                  111 Eight Avenue (JFK10)    New York    NY 1.37         6             8           93       0.046686
          4                       The Westin Building     Seattle    WA 1.75        10             0           44      -0.277665
          4         43830 Devin Shafron Drive (IAD24)     Ashburn    VA 1.15        21             0           14       0.286869
          4       43940 Digital Loudoun Plaza (IAD35)     Ashburn    VA 1.15        21             0           12       0.288889
          4                           NTT Ashburn VA1     Ashburn    VA 1.15        21             0           25       0.275758
          4                            Aligned IAD-02     Ashburn    VA 1.15        21             0            5       0.295960
          4              Crown Castle Los Angeles LA2 Los Angeles    CA 1.22         1             8           20       0.137731
```

Figure A13: NSGA-II Optimization Results

Top-ranked data centers balance PUE, IXP, service features, and age for optimal trade-offs.

Key outcomes included:

- **Best Energy Efficiency (Lowest PUE):** 1.15
- **Maximum Connectivity (Highest IXP Count):** 25
- **Top Infrastructure Availability:** Avg. 9.5/10
- **Optimal Facility Age Range:** Between 5 and 12 years

Note: Here's the link to the data set we had scrapped and built the model on Dataset .

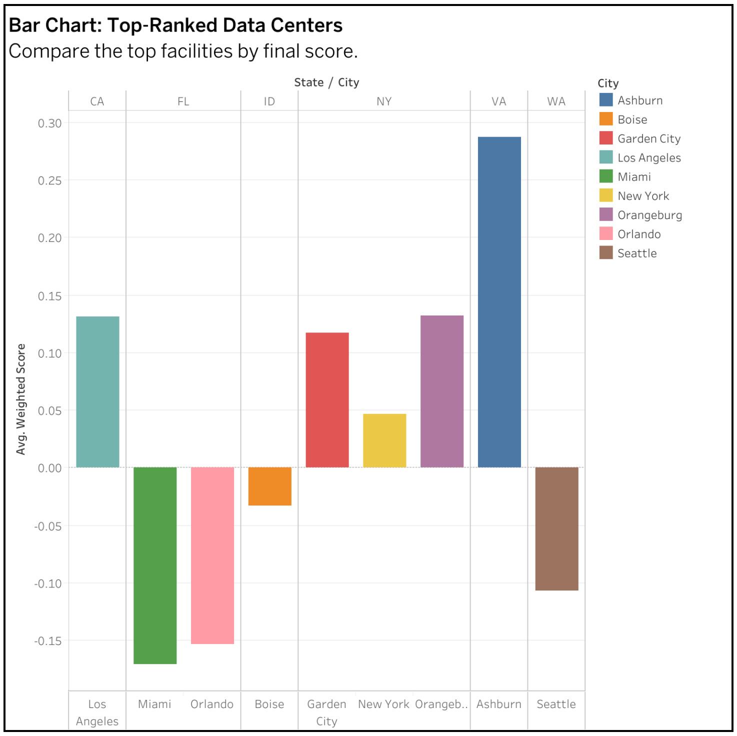


Figure A14: Top-Ranked Data Centers by Score

Ashburn leads in average score; Miami and Seattle score lowest among top data center cities.

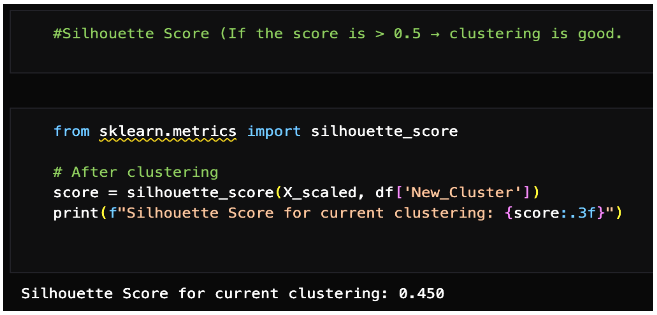
The above is an illustration of one example of the solution generated but general the goal here is to not give one optimal solution rather a series of most optimal solutions which stakeholders can choose from based on their trade-off choices.

5. Analytical Results and Interpretations

The project yielded the following key insights:

5.1 Clustering & PCA:

To check if the cluster classification is accurate or not we calculated it with a Silhouette Score.



```
#Silhouette Score (If the score is > 0.5 → clustering is good.
```

```
from sklearn.metrics import silhouette_score

# After clustering
score = silhouette_score(X_scaled, df['New_Cluster'])
print(f"Silhouette Score for current clustering: {score:.3f}")
```

```
Silhouette Score for current clustering: 0.450
```

Figure A15: Silhouette Score for Cluster Validation
To evaluate clustering quality, a **Silhouette Score** was computed using sklearn.metrics.silhouette_score.
The score ranges from -1 to 1, where:

Note: **> 0.5** indicates well-separated, high-quality clusters.
0.45 (current result) suggests **moderately good clustering**—some overlap, but overall meaningful structure.
This helps confirm that the cluster classification has reasonable internal cohesion and external separation.

Three distinct operational clusters were identified and Cluster Profiles were created for classification and informed decision making :

- **Cluster 0:** Medium-sized, efficient with balanced infrastructure.
- **Cluster 1:** Large-scale, minimal services, high efficiency.
- **Cluster 2:** Small, service-rich, lower efficiency (high PUE).

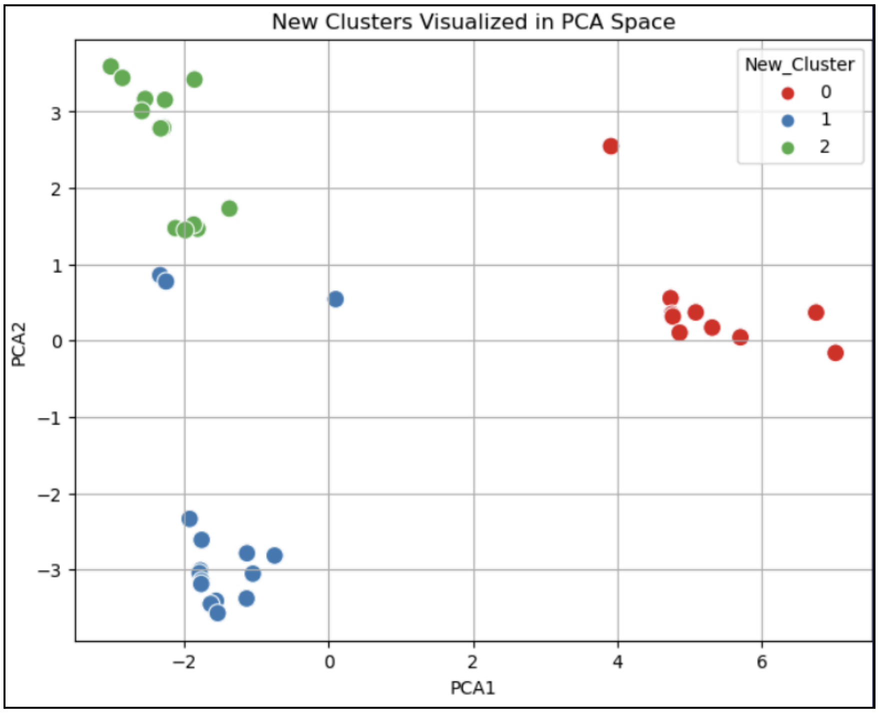


Figure A16: PCA-Based Cluster Visualization
PCA shows clear spatial separation of clusters, visually confirming clustering structure and supporting model validity.

5.2 Gravity Model: **Gravity Model Results (India Data Set)**

Purpose: Rank Indian cities for potential new data centers.

Action Taken:

- Used weighted metrics: ENERGY cost, LAND cost, IXPs, Workforce, Climate, Renewable %, Availability.

- **Top Ranked Cities:**
  - **Hyderabad** – Score: 0.783
    Best energy rate (6.3 ₹/kWh), high land availability, strong network (IXP index = 8).
  - **Bangalore** – Score: 0.742
    Strong sustainability (renewables 70.8%), connectivity.
  - **Pune** – Score: 0.708
    Balanced metrics across cost, scalability, and environment.

Value for Business: Guides stakeholders with optimal trade-offs in terms of cost, performance, availability and scalability.

Note: We have explained the model in depth above with all the objective function, Constraints, Data Source and Weights

5.3 NSGA-II Optimization:

- PUE value between 1.15
- IXP count up to 25
- Infrastructure Score ≥ 9.5
- Age <= 10 years

These above are general optimal trade-offs that help compare sites objectively but also dependent on business objectives to prioritize one over the other. The above are based on the dataset here. The data centers from these cities were considered:

Value for Business: Guides stakeholders with optimal trade-offs in terms of energy efficiency (PUE), Internet Connectivity (IXP), Infrastructure (Service Availability Score) and Scalability (Age of Data Center).

Note: The data centers from these cities were considered**,** Miami, Orlando, Boise, New

York, Orangeburg, Garden City, Seattle, Ashburn and Los Angeles

6. Conclusions and Business Impact

This project provided a comprehensive, data-driven framework to optimize data center planning and operations:

- **Strategic Clustering** enabled segmentation of existing centers into meaningful types, guiding targeted upgrades, consolidation, or expansion.

- **Predictive Planning** allowed accurate classification (100% accuracy) of future data centers into operational clusters, reducing guesswork in planning.

- **Gravity Model** prioritized Indian cities for expansion, identifying Hyderabad, Bangalore, and Pune as top locations based on cost, connectivity, and scalability.

- **NSGA-II Optimization** offered balanced solutions across efficiency, connectivity, and infrastructure readiness.

Business Impact:

- Supports faster, smarter decision-making with quantifiable benchmarks.
- Reduces capital and operational risks by aligning site selection with business objectives.
- Enables proactive capacity planning and better alignment with sustainability goals.

Overall, the analysis bridges technical evaluation with business strategy, making data center investments more efficient, scalable, and future-ready.

7. Contributions

Mahima Advilkar: Led the Project, clustering and exploratory data analysis (EDA), built the classification model for informed decision making, and provided interpretation of analytical outcomes and worked on presentation and report.

Prasad Pilankar: Designed and implemented the Gravity Model, including criteria weighting and city-level ranking framework.

Abhinav Krishnamoorthy: Collected and cleaned datasets, executed NSGA-II multi-objective optimization, and developed presentation and report.

Navin Kishore Ravi Kumar: Synthesized business insights, structured the report, and aligned analytical outputs with strategic objectives.

8. References

1. K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, “A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II,” *IEEE Transactions on Evolutionary Computation*, vol. 6, no. 2, pp. 182–197, Apr. 2002. Available: https://ieeexplore.ieee.org/document/996017

2. Buyya, R., Srirama, S. N., Casale, G., Calheiros, R. N., & Simmhan, Y. (2018). “A Manifesto for Future Generation Cloud Computing: Research Directions for the Next Decade,” ACM Computing Surveys, 51(5), 1–38. Available at: https://arxiv.org/abs/1711.09123

3. Jin, X., Zhang, F., Vasilakos, A., & Liu, Z. (2016). Green Data Centers: A Survey, Perspectives, and Future Directions. *ResearchGate*. Link: https://www.researchgate.net/publication/305779625_Green_Data_Centers_A_Survey_Perspectives_and_Future_Directions

4. Amazon Web Services, “What is a Data Center?” [Online]. Available: https://aws.amazon.com/what-is/data-center/

5. IBM, “What is a Data Center?” [Online]. Available: https://www.ibm.com/think/topics/data-centers

6. International Energy Agency (IEA), “Electricity consumption of data centres worldwide,” 2023. [Online]. Available:https://www.iea.org/energy-system/buildings/data-centres-and-data-transmission-networks

7. DEAP - Distributed Evolutionary Algorithms in Python, “NSGA-II Documentation.” [Online]. Available: https://deap.readthedocs.io

8.DataCenterMap, “Global Data Center Listings.” [Online]. Available: https://www.datacentermap.com

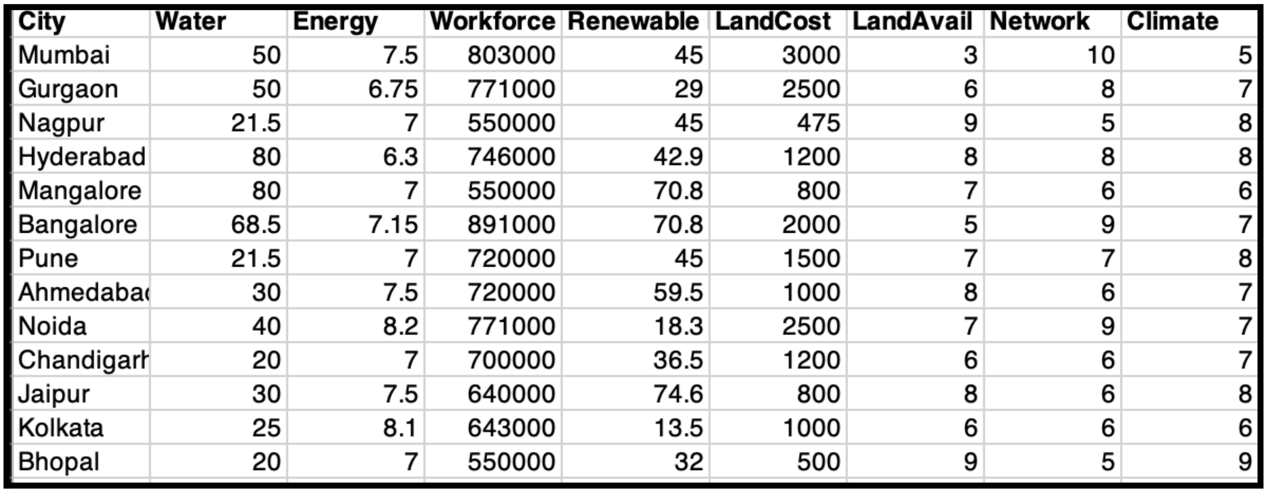
9. TierPoint, “10 Data Center Location Strategies to Consider,” 2023. [Online]. Available: https://www.tierpoint.com/blog/data-center-location-strategies

10. Liu, B., Xu, W., Zhang, X., & Wu, J. (2024). “Multi-Objective Optimization Study on Data Center Location and Resource Allocation within China's 'East Data West Computing' Strategy.” *SSRN*. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5057250

11. Zhang, T., & Liu, W. (2024). *Data-guided Gravity Model for Competitive Facility Location*. ResearchGate. https://www.researchgate.net/publication/378189001

---

9. Appendix (Code Snippets & Tables)



City	Water	Energy	Workforce	Renewable	LandCost	LandAvail	Network	Climate
Mumbai	50	7.5	803000	45	3000	3	10	5
Gurgaon	50	6.75	771000	29	2500	6	8	7
Nagpur	21.5	7	550000	45	475	9	5	8
Hyderabad	80	6.3	746000	42.9	1200	8	8	8
Mangalore	80	7	550000	70.8	800	7	6	6
Bangalore	68.5	7.15	891000	70.8	2000	5	9	7
Pune	21.5	7	720000	45	1500	7	7	8
Ahmedabad	30	7.5	720000	59.5	1000	8	6	7
Noida	40	8.2	771000	18.3	2500	7	9	7
Chandigarh	20	7	700000	36.5	1200	6	6	7
Jaipur	30	7.5	640000	74.6	800	8	6	8
Kolkata	25	8.1	643000	13.5	1000	6	6	6
Bhopal	20	7	550000	32	500	9	5	9

Table A2: Data Used for Gravity Model

9.1 Clustering and PCA

```
Included critical analytical code snippets for clustering, PCA visualization, predictive
modeling, gravity modeling, and NSGA-II optimization.

from sklearn.cluster import KMeans
from sklearn.decomposition import PCA
from sklearn.preprocessing import StandardScaler
```

```
import matplotlib.pyplot as plt

# Standardizing the features
scaler = StandardScaler()
X_scaled = scaler.fit_transform(data)

# PCA for dimensionality reduction
pca = PCA(n_components=2)
X_pca = pca.fit_transform(X_scaled)

# KMeans clustering
kmeans = KMeans(n_clusters=3, random_state=42)
clusters = kmeans.fit_predict(X_pca)

# Plotting PCA results
plt.scatter(X_pca[:,0], X_pca[:,1], c=clusters, cmap='viridis', alpha=0.7)
plt.xlabel('PCA Component 1')
plt.ylabel('PCA Component 2')
plt.title('Data Center Clusters via PCA')
plt.show()
```

9.2 Gravity Model

```
import pandas as pd

# Calculate attraction score based on criteria weights
weights = {'Cost': 0.3, 'Connectivity': 0.25, 'Renewable': 0.15, 'Workforce': 0.1, 'Land': 0.1, 'Climate': 0.1}
data['Gravity_Score'] = (
    data['Cost'] * weights['Cost'] +
    data['Connectivity'] * weights['Connectivity'] +
    data['Renewable'] * weights['Renewable'] +
    data['Workforce'] * weights['Workforce'] +
    data['Land'] * weights['Land'] +
    data['Climate'] * weights['Climate']
)
data_sorted = data.sort_values(by='Gravity_Score', ascending=False)
print(data_sorted[['City', 'Gravity_Score']].head())
```

9.3 NSGA-II Multi-Objective Optimization

```
from deap import base, creator, tools, algorithms
import random

# Defining objectives (minimize PUE, maximize connectivity and robustness, minimize
age)
creator.create("FitnessMulti", base.Fitness, weights=(-1.0, 1.0, 1.0, -1.0))
creator.create("Individual", list, fitness=creator.FitnessMulti)

# Toolbox setup
toolbox = base.Toolbox()
toolbox.register("attr_float", random.uniform, 0, 1)
toolbox.register("individual", tools.initRepeat, creator.Individual, toolbox.attr_float,
n=4)
toolbox.register("population", tools.initRepeat, list, toolbox.individual)

# Objective function
def evalDataCenter(individual):
    pue, connectivity, robustness, age = individual
    return pue, connectivity, robustness, age

toolbox.register("evaluate", evalDataCenter)
toolbox.register("mate", tools.cxSimulatedBinaryBounded, low=0, up=1, eta=20.0)
toolbox.register("mutate", tools.mutPolynomialBounded, low=0, up=1, eta=20.0,
indpb=0.2)
toolbox.register("select", tools.selNSGA2)

# Running NSGA-II
pop = toolbox.population(n=100)
algorithms.eaMuPlusLambda(pop, toolbox, mu=100, lambda_=200, cxpb=0.7,
mutpb=0.3, ngen=50, stats=None)

# Extracting Pareto optimal solutions
pareto_front = tools.sortNondominated(pop, len(pop), first_front_only=True)[0]
for ind in pareto_front:
    print(ind, ind.fitness.values)
```