Nil Beserler
01/10/2024

Telecommunications Disaster Inquiry, Accenture
Fall 2023, AI Studio Project Write Up

Business Focus

The goal of the project is to utilize unsupervised learning techniques to analyze and predict the impact of natural disasters on cellular towers throughout the United States. By identifying useful clusters within the data, the project aims to create estimations that can be visualized to uncover critical insights. These insights will enable the telecommunications client to make informed decisions about disaster preparedness and infrastructure resilience to mitigate the effects of natural disasters on their network.

Data Preparation and Validation

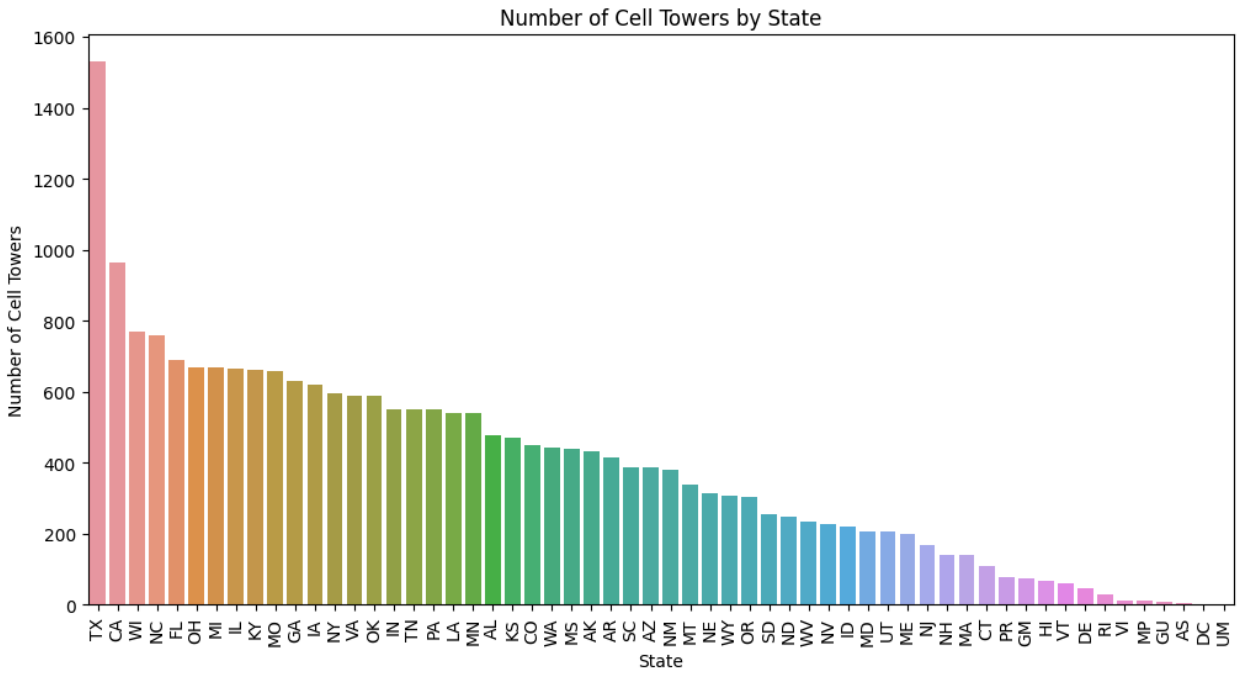
Dataset Description

We were given two datasets by Accenture, a disasters declaration dataset and a cellular towers dataset. Disaster declaration datatset included information like incident type, deceleration type, and years of disasters. Whereas the cellular towers dataset included features such as tower counts, addresses, cities, and counties.

Exploratory Data Analysis

Data exploration in this project was a challenging task. We started by removing duplicate rows and addressing missing data where possible, standardizing variables, and fixing inconsistencies in both datasets.

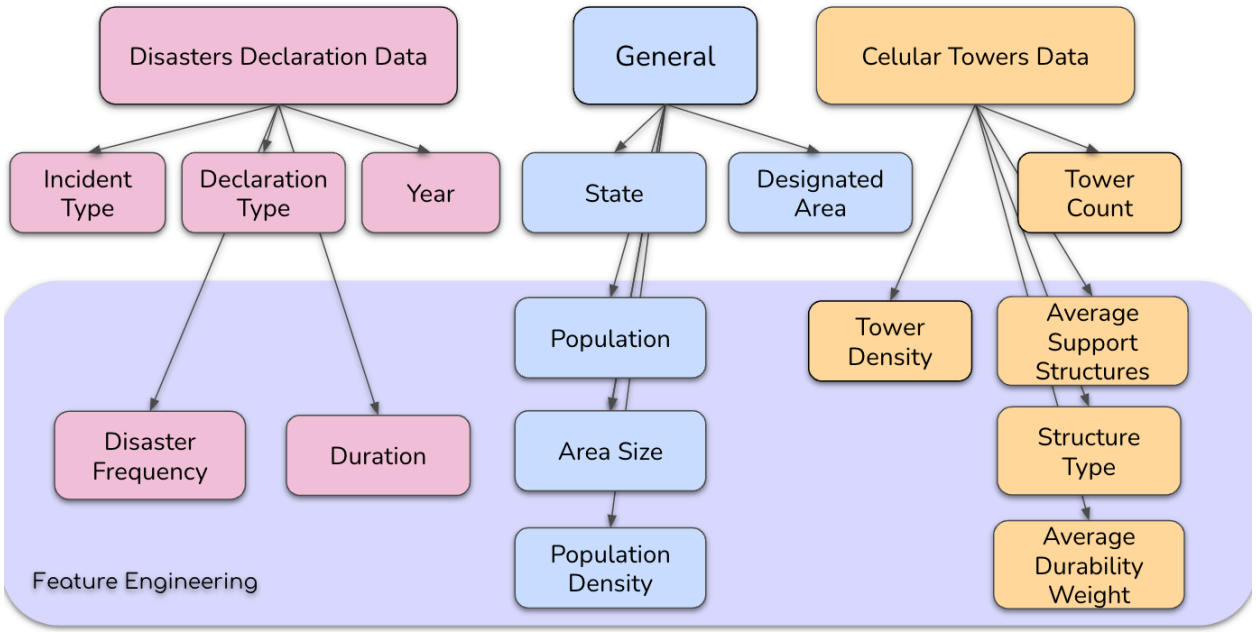
Next, we turned our attention to the cellular tower dataset. After examining its content and statistics, we added another cellular tower dataset to it. This step helped deal with the skewed nature of the original data. Although there was still some skewness due to the nature of cellular towers, our efforts led to a more updated and representative dataset.



Then, we focused on the disaster declaration dataset. During this phase, we cleaned the data again, which included renaming columns for better understanding and removing irrelevant columns like 'fiscal year declared' and 'declaration request number'. We also simplified some features, for example, changing 'start' and 'end' dates to a single 'duration' feature. An important part of this process was manually aligning states and counties to match those in the cellular tower dataset, ensuring the two datasets were consistent and compatible.

Feature Engineering and Selection

One challenge we faced with our datasets was their lack of critical information for decision-making about cellular tower needs and maintenance. It became evident that relying solely on disaster declarations and cell tower locations was insufficient because such data does not account for the nuanced impact of disasters on varying population densities and infrastructure resilience. To address this, we expanded our dataset to include variables essential for informed decision-making.



We began by enhancing our original disaster declarations dataset through feature engineering, adding key metrics like disaster frequency and duration. Furthermore, we integrated an additional dataset providing crucial details like population, area size, and population density for each area.

For a more comprehensive analysis, we incorporated "Structure Type," categorizing cellular towers into types such as lattice or monopole, each with distinct vulnerabilities to natural disasters. This involved extensive research and the assignment of specific weights to each tower type based on their resilience. Similarly, "Durability Weight" was introduced to quantify the robustness of various cellular structures, with higher values indicating an enhanced ability to withstand natural disasters. The feature "Support Structure" was also added to address the elements that provide stability and strength to the main structure.

Moreover, we calculated tower density by considering the location and count of towers in each area. By adding these variables, we have significantly improved our dataset's capability. Now, it offers a more detailed and effective tool for our clients, aiding them in making informed decisions about disaster preparedness and the maintenance of cellular infrastructure.

Merging and Preparing for Model Training

In the final stage of dataset preparation, we merged the two primary datasets. A key challenge we faced was the differing granularity levels; the cellular tower dataset detailed each tower by county, state, and coordinates, whereas the disaster dataset sometimes listed events as

"Statewide" or specific to certain counties. Including more detail was important to maintain the specificity of the data, particularly for localized analysis. To address this and retain as much detail as possible, we merged based on state and county, labeling statewide disasters accordingly and including only state information for them.

Following this, we encoded the categorical variables using one-hot encoding and standardized the numerical variables. This step was essential for the unsupervised machine learning models we planned to use.

Approach

While considering the choice of model, we decided that a clustering algorithm would be suitable for grouping similar data points based on various attributes. The use of a supervised learning model was not feasible, as our dataset lacked labels. Consequently, we decided to start our modeling with a simple k-means algorithm to observe how clustering performs on our dataset. Following this, we planned to implement a more advanced clustering model; the DBSCAN model, which is better suited to capture the true complexity of our dataset.

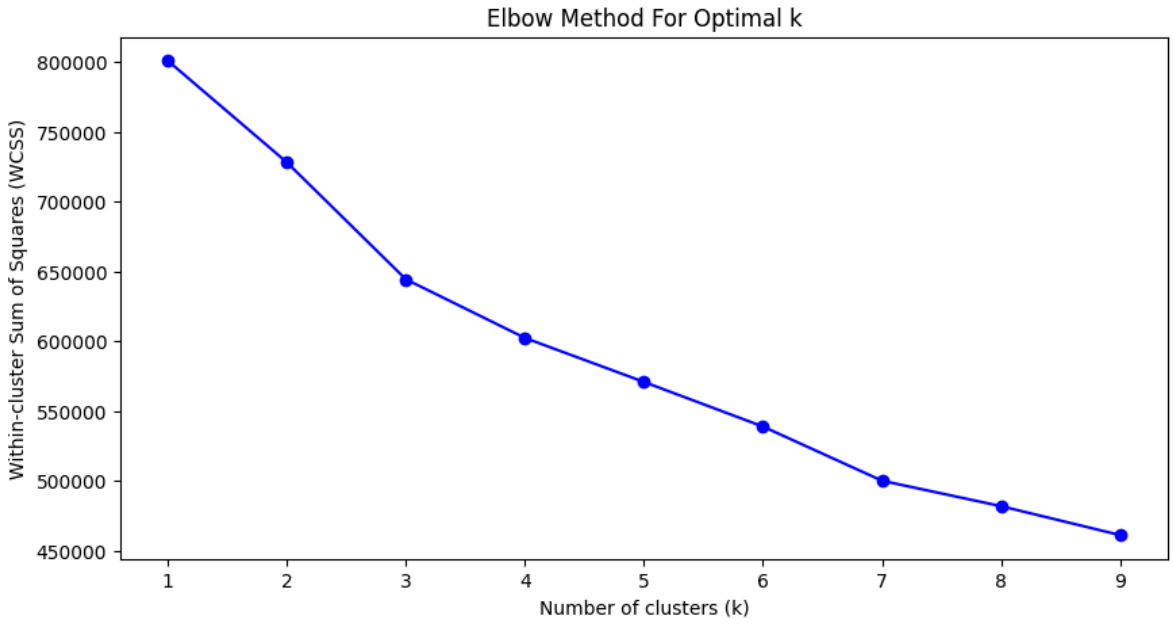
PCA

Our dataset expanded to 1938 columns due to the encoded features, making it computationally demanding to manage. To address this, we applied Principal Component Analysis (PCA), which effectively reduced the number of components to 37 while still explaining 90% of the variance. This reduction was key in preserving the nuances of our dataset while significantly lowering the computational resources required for processing.

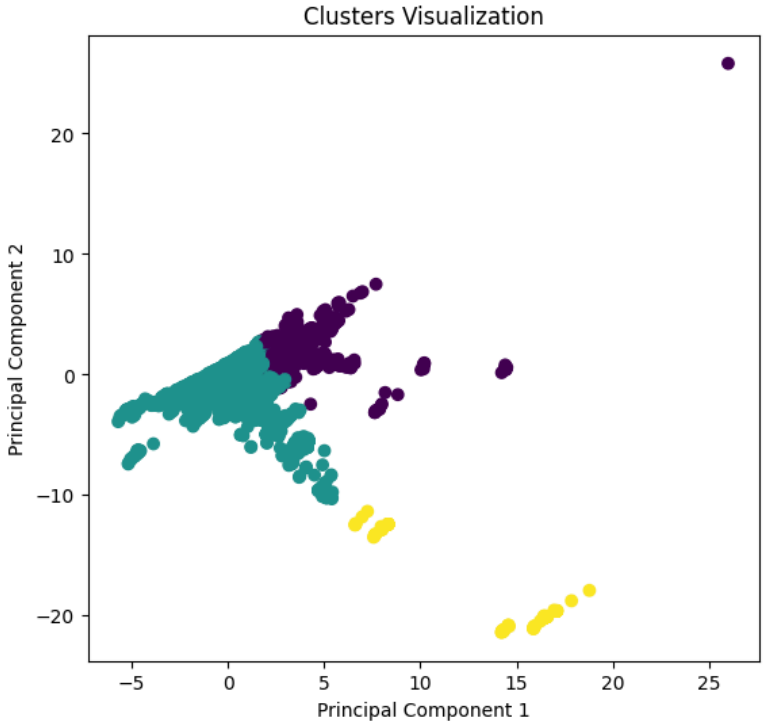
K-Means

We started by using the elbow method to figure out how many clusters we needed. Since there wasn’t a clear "elbow" showing up, we tried different K values and looked at the Silhouette Score for each. The Silhouette Score goes from -1 to 1, where a high score means the data fits well in its own cluster and not so much in the neighboring ones. We used this score to see how well our models were making clusters. After checking out different scores, we decided that using

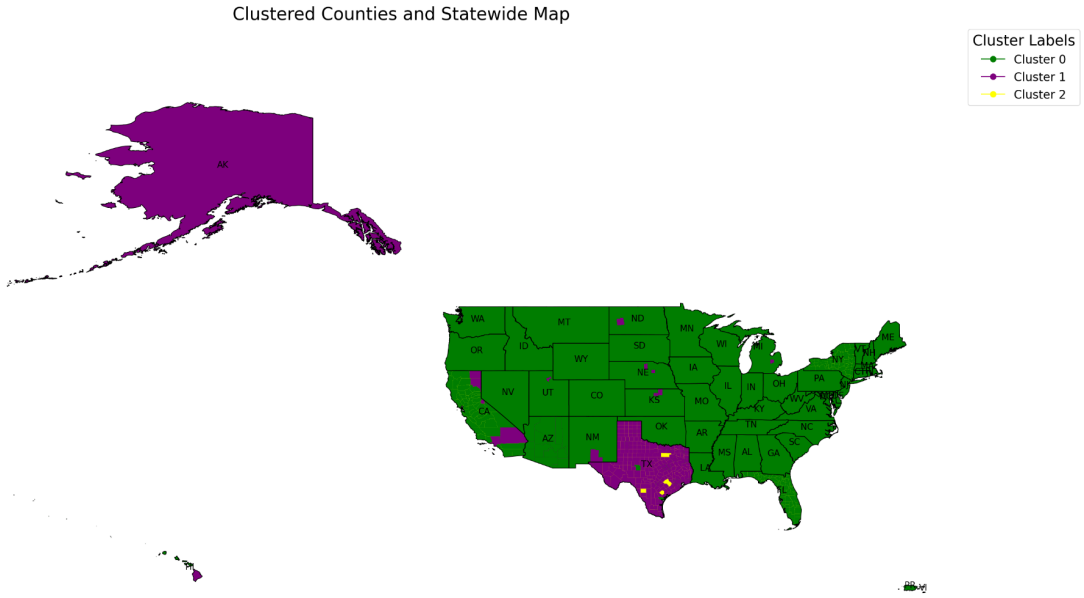
a K value of 3 was the best choice for our data.



Training the k-means model and then visualizing our results showed that our data formed three distinct clusters of varying sizes, with similar data points grouped together. The visualization was plotted on principal component axes. The y-axis reflected physical characteristics like durability, tower count, and substructures. On the other hand, the x-axis (the second component) captured variations in temporal and geographical factors, including year, duration, and area size. This helped us understand how different features influenced the clustering.



By analyzing the statistical trends for each cluster, and then highlighting the clusters on a geographical map we were able to draw meaningful insights



Cluster 0 is the largest, comprising 48,805 samples. It's characterized by a higher durability, suggesting it is made up of robust infrastructure, but it has a lower tower count, which may indicate these are rural areas. The duration of disasters in this cluster is average, with a moderate risk assessment. The investment priority for this cluster is to maintain the current level of investment.

Cluster 1, a medium-sized cluster with 4,490 samples, shows lower durability and below-average tower count. It stands out for having longer durations of disasters, leading to a high-risk assessment. Therefore, the recommendation for this cluster is to increase investment, likely to improve infrastructure resilience.

The smallest cluster, Cluster 2, with only 147 samples, has slightly below-average durability and a significantly higher tower count. Disasters in this cluster tend to have below-average durations. The risk assessment is moderate, but like Cluster 1, the suggestion is to increase investment in resilience measures.

These findings can be summarized as:

Features \ Clusters	0	1	2
Size	48805 Samples Largest Cluster	4490 Samples Medium Sized	147 Samples Smallest Cluster
Durability	Higher durability suggesting robust infrastructure	Lower Durability	Slightly below average
Tower Count	Lower count, may indicate rural areas	Below Average	Significantly high tower count
Duration of Disasters	Average duration	Longer Durations	Below Average Duration
Risk Assessment	Moderate risk	High risk	Moderate risk
Investment Priority	Maintain current investment	Increase investment	Increase investment for resilience measures

DBSCAN

Transitioning from the k-means clustering, we considered the density-based spatial clustering of applications with noise (DBSCAN), which offered several advantages for our cellular tower dataset. DBSCAN excels in environments where the density of towers varies significantly, such as between urban and rural regions. It can adapt to these density variations by grouping towers based on proximity and density criteria. This feature is particularly useful for our dataset, where we expect the density of towers to be inconsistent across different areas.

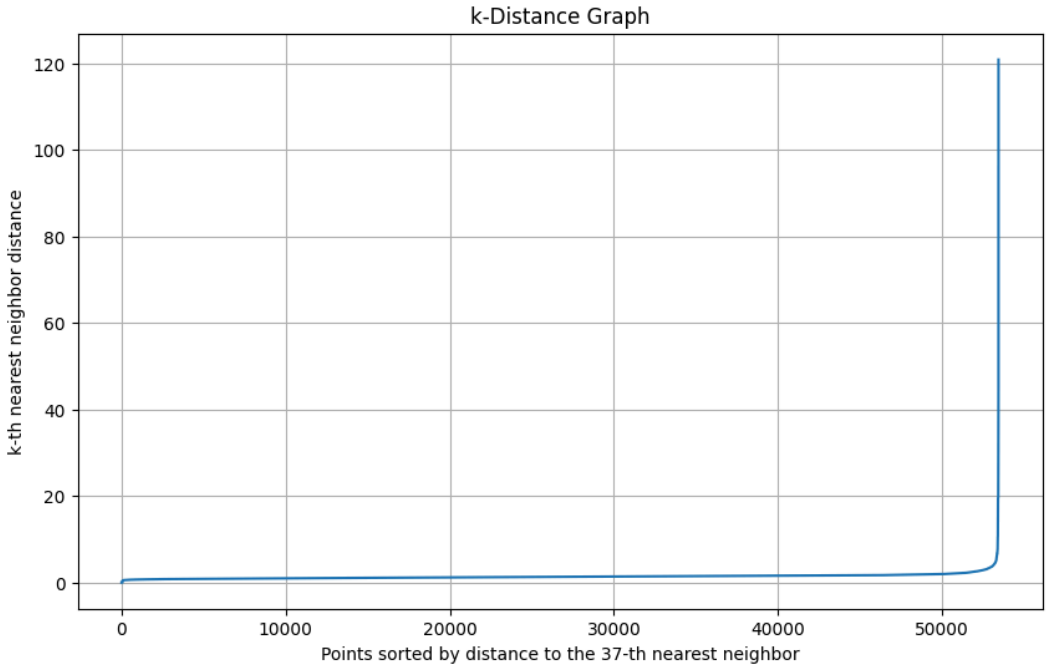
Moreover, DBSCAN's ability to handle noise effectively allows it to identify isolated towers or outliers that do not fit into the common patterns. This characteristic is significant when dealing with real-world data that often contains anomalies. The algorithm's capacity to automatically determine the number of clusters is beneficial when dealing with datasets that have groups of vastly different sizes, as is common with cellular towers where some regions may have many towers while others have few.

Lastly, DBSCAN's versatility with various data distributions makes it a robust choice for our cellular tower data, ensuring that both densely clustered urban areas and sparsely populated rural regions are accurately represented in the analysis. The transition to DBSCAN, therefore, seemed a natural step in our efforts to capture the complex patterns present in our dataset.

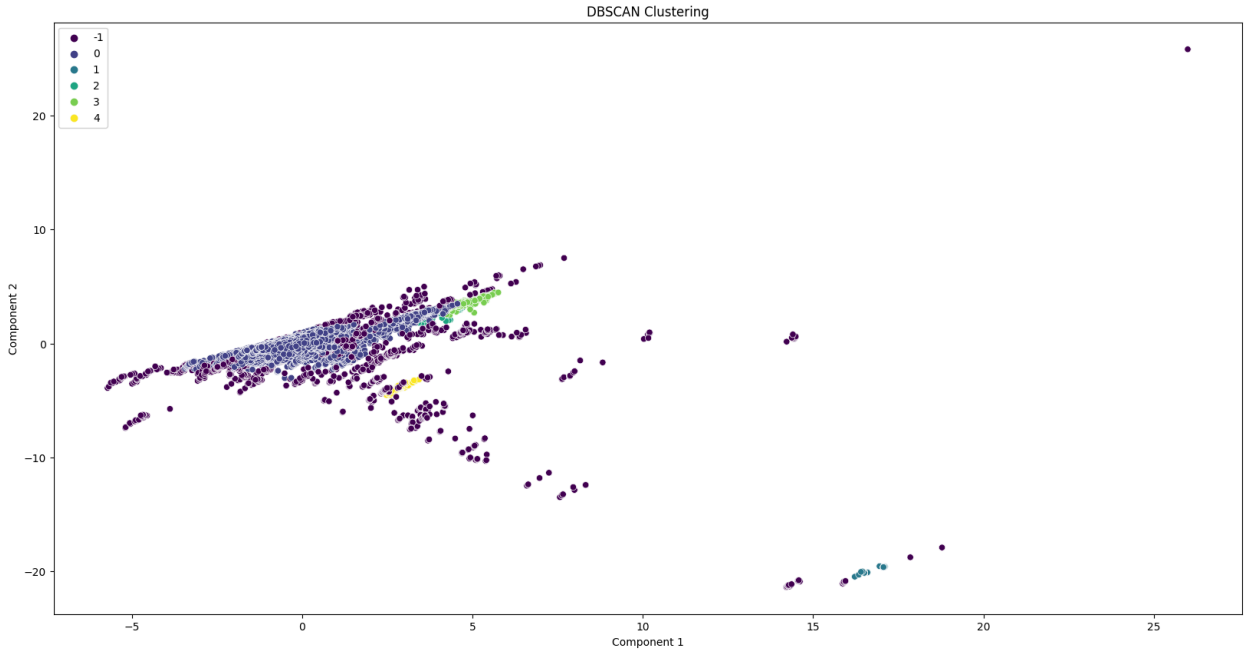
In our DBSCAN model, the parameters of minimum samples (min samples) and epsilon needed to be finetuned. Min samples, denoted as 'k', is crucial because it sets the minimum number of neighboring towers needed for a given tower to be considered a core member of a cluster. This essentially determines the density threshold for cluster formation. We decided on min samples by testing out different values and then deciding on using the default value for computational efficiency.

Epsilon (ε), on the other hand, determines the maximum distance between two towers for them to be classified as part of the same cluster. It's a measure of proximity and a key factor in defining the spatial reach of each cluster.

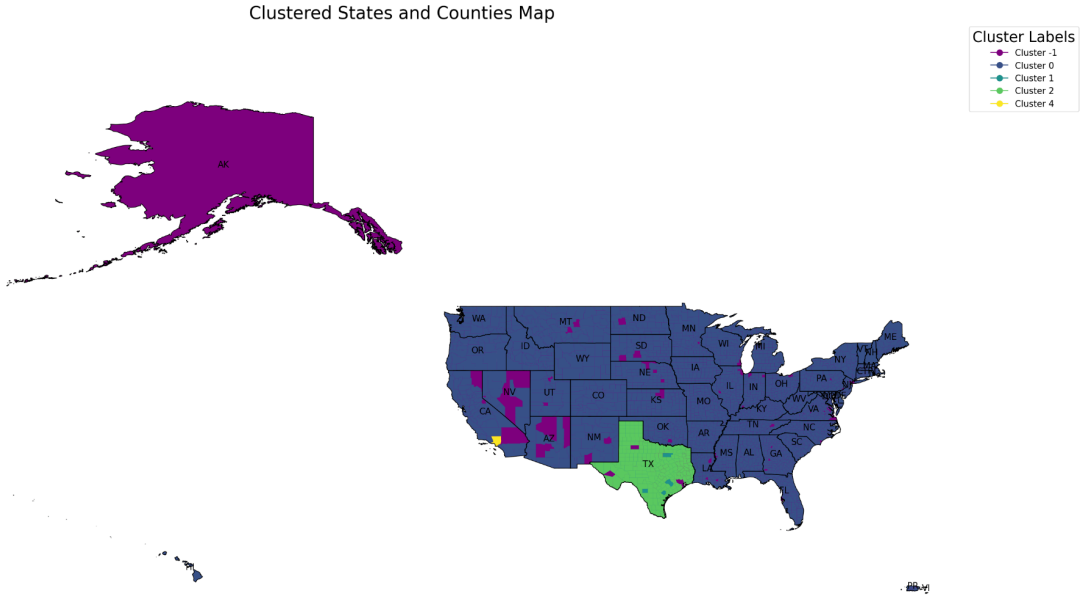
To fine-tune the epsilon parameter effectively, we used the K distance graph. This graph plots the distance to the k-th nearest neighbor for each point, allowing us to visually assess where the distances start to increase significantly. This point, often referred to as the "elbow," is where the epsilon value is ideally set. For our model, an epsilon value of 2 emerged as the optimal choice. Setting epsilon at this level ensures that our clusters are both meaningful and reflective of the actual patterns in the tower distribution data, capturing the nuances of how towers are spaced across different regions.



After running and visualizing our DBSCAN we got the following results:



Similar to our KMeans model we mapped these clusters on a geographical map as:



Cluster -1, with 1,623 noise samples, has been assessed as a moderate risk. Despite its below-average durability and above-average tower count, it suggests that standard maintenance is adequate for this group.

Cluster 0 is the largest, encompassing 50,699 samples, and is also deemed a moderate risk. It has average durability and a below-average tower count, with slightly below-average disaster durations. Similar to Cluster -1, standard maintenance is recommended for this cluster.

Cluster 1, though small with just 90 samples, is considered high risk due to its significantly below-average durability and extremely high tower count. This cluster's particular vulnerability necessitates targeted investment to bolster its resilience.

Clusters 2 and 3, with 968 samples combined, present a moderate risk. The data suggests that while the towers might be newer, their ability to withstand prolonged disasters could be lacking, indicating a need for strategic investment to enhance durability.

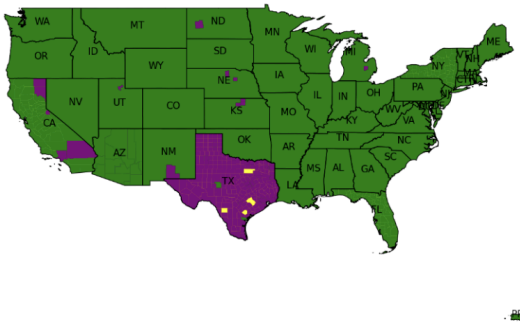
Lastly, Cluster 4, although small with only 62 samples, is labeled high risk because of its below-average durability. The particular needs of this cluster, likely influenced by its topography, require targeted investment, particularly to increase cell tower density in preparation for severe disasters.

These findings can be summarized as follows:

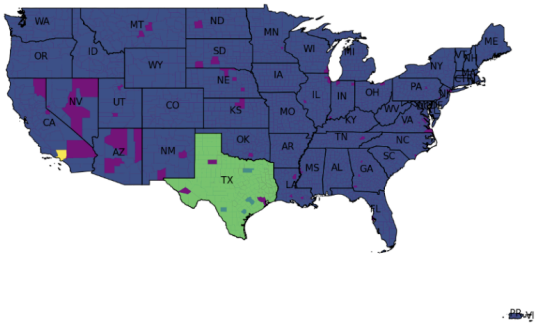
Clusters / Features	-1	0	1	2 & 3	4
Size	1623 Noise samples	50,699 Largest cluster	90 samples	968 samples	62 samples
Durability	Below average	Average durability	Significantly below average	Significantly below average	Below average
Tower Count	Above average	Below average	Extremely high count	Below average	Reasonable average
Duration of Disasters	Slightly below average	Slightly below average	Below average duration	Above average	Significantly below average
Risk Assessment	Moderate risk	Moderate Risk due to large size	High Risk	Moderate Risk	High risk
Investment Priority	Standard maintenance	Standard maintenance	Targeted Investment	Evaluate for strategic investment	Targeted investment

Key Findings and Insights

K-means



DBScan



Upon comparing the maps generated by our models, we observed that the largest clusters were consistently concentrated in the same specific regions across both maps. These maps reveal that these prominent clusters share distinct characteristics: they are located in areas with a lower propensity for disasters, possess a sparse density of cell towers, and exhibit a lower durability weight. Such insights are invaluable for guiding infrastructure investment and maintenance strategies. They suggest the potential necessity of reinforcing existing towers or considering the construction of new ones in these particular areas, ensuring more robust and resilient cellular networks.

In our analysis, we utilized two distinct clustering algorithms to organize our dataset. The first, K-Means Clustering, partitions the dataset into K distinct clusters by calculating the mean of the data points in each cluster. This model produced three clusters with a Silhouette Score of 0.3811, indicating a reasonable level of separation between clusters. Its advantages include ease of implementation, computational efficiency, particularly for large datasets, and the ability to create tighter clusters compared to hierarchical clustering. However, its limitations are significant; the number of clusters must be predetermined, it is sensitive to the initial placement of centroids, and it operates under the assumption that clusters are spherical and uniformly sized, which is often not the case with real-world data.

The second model, DBSCAN (Density-Based Spatial Clustering of Applications with Noise), takes a different approach by identifying clusters as areas of high density separated by areas of low density, which are labeled as outliers. It found five clusters and achieved a Silhouette Score of 0.4707, suggesting better cluster definition compared to K-Means. DBSCAN's strengths lie in its ability to detect clusters of arbitrary shape and its robustness to outliers. It does not require a predefined number of clusters, which can be considered an advantage over K-Means. However, DBSCAN also has its drawbacks, such as the complexity of choosing appropriate values for its

two parameters, epsilon (eps) and minimum samples (min_samples). It is also less effective with high-dimensional data and can struggle with clusters of varying densities.

Overall, while both models have their pros and cons, the Silhouette Scores indicate that DBSCAN slightly outperformed K-Means in our context, providing a more defined clustering solution.

Learning Outcomes and Contributions

Through this project, I gained valuable insights into the practical applications of data-driven decision-making. By employing clustering techniques, I improved my unsupervised learning skills and effectively visualized and communicated the insights I gathered. The data preparation process reinforced the importance of data quality and its impact on model performance, reminding me not to underestimate it. Additionally, I learned to more accurately estimate the time required for each project phase and to plan accordingly.

The presentation process highlighted the significance of maintaining a balance between technical/programming end and business perspectives. I learned that understanding the data in the models is essential, but equally important is the ability to translate these into actionable business strategies. Overall, my experience with cluster analysis has been a comprehensive lesson in integrating data science with strategic business planning for informed, predictive decision-making.

I would like to express my deepest gratitude to my teammate, Stellar Nguyen, for her dedication and support throughout the project. Her hard work was indispensable. I am also thankful to our tutor, Kaifeng Pang, and our challenge advisors, Shreya Sheth and Khabab Salama, who were always ready to assist us at a moment's notice whenever we encountered questions. A special thanks to the Breakthrough Tech AI team for organizing and steering us through this valuable learning experience. Their guidance was fundamental to our success.

Github Repo:

https://github.com/NilBeserler/Telecommunications_Disaster_Inquiry_Accenture