Rensselaer Polytechnic Institute

Identifying Vulnerable Child Care Centers Due to Effects of Temperature and Precipitation

Kerui Wu wuk9@rpi.edu

Weihao Li liw21@rpi.edu

Yanfeng Liu liuy63@rpi.edu

Chiting Lu luc6@rpi.edu

Jinhui Yu jinhuy@rpi.edu

Tianze Zhu zhut3@rpi.edu

Data Science - ITWS 6350

Thilanka Munasinghe

December 7, 2023

Abstract

Childcare centers gather lots of children who need protection. Focusing on minimizing physical harm, the impact of natural hazards induced by precipitation and temperature fluctuations on childcare center buildings is a critical concern. Heavy rainfall and extreme temperatures can compromise structural integrity and pose health risks. Leveraging historical data from NASA's Global Precipitation Mission and childcare center locations from Homeland Infrastructure Foundation-Level Data, we processed and cleaned the dataset, utilizing machine learning clustering algorithms. The K-means clustering algorithm and Gaussian Mixture Model, among others, classified childcare centers based on precipitation and temperature data, with silhouette scores above 0.3. Visualization on a US map revealed varying risks among states, influenced by environmental factors. This study underscores the importance of incorporating additional elements like wind and snowfall in future research to assess childcare facility risks and regional vulnerabilities comprehensively.

Keywords: Child Care, Temperature, Rainfall, Clustering, Machine Learning

1 Introduction

Childcare licensing [1] is a regulatory process implemented by government authorities at the state or territory level to establish and enforce specific standards for the operation of childcare programs. While the primary objective of child care licensing is to safeguard the health, safety, and well-being of children in care, the foundation of child protection lies in minimizing the risk of physical harm [2]. This involves ensuring that buildings are structurally sound and free from natural hazards, adhering to proper food service and sanitation practices, and having program policies that equip staff to respond effectively to emergencies.

Among those factors, the impact of natural hazards, specifically those induced by precipitation and temperature fluctuations, on childcare center buildings is a critical concern. Changes in weather patterns, such as heavy rainfall, storms, or extreme temperatures, can pose significant risks to childcare facilities' structural integrity and safety. Excessive precipitation may lead to flooding or water damage, compromising the building's foundation and creating potential hazards. On the other hand, extreme temperatures, whether excessively hot or cold, can strain a facility's infrastructure, affecting heating, ventilation, and air conditioning systems. Such conditions not only jeopardize the physical safety of the building but also pose potential health risks to the children and staff within. Therefore, addressing and mitigating the impact of natural hazards caused by precipitation and temperature is paramount in ensuring the overall safety and resilience of childcare centers.

Historical precipitation and surface temperature records offer a reliable estimate of regions prone to heightened risk. This information can be instrumental in addressing and mitigating potential infrastructure losses in those areas. Therefore, We took advantage of the precipitation and surface temperature data obtained from the NASA Global Precipitation

Mission (GPM) and child care center locations from Homeland Infrastructure Foundation-Level Data (HIFLD) to classify vulnerable child care centers in Florida, US.

We filtered the dataset from HIFLD to get every childcare center building's longitude and latitude, matched these coordinates with the earth's rainfall and temperature dataset from 03/01/2013 to 03/01/2023, and removed extreme values(random noise) to get the training set we used in this project.

We implemented multiple Machine Learning clustering algorithms including centroid-based algorithms, density-based algorithms, and spectral clustering algorithms to classify the childcare center based on the processed data, and evaluated our clustering result by calculating the silhouette score, a metric that measures the clustering quality in a dataset. It turned out that every algorithm's score was above 0.3. More than that, the centroid-based algorithm obtained 0.56, indicating that the clusters are well separated.

By visualizing our clusters on a map, we've identified varying risks among states, influenced by the distribution of precipitation and surface temperature throughout different months in a given year. This suggests that additional environmental factors will also be significant in assessing the risk associated with a childcare facility and the overall region. These factors will serve as the foundation for our forthcoming research, which is to combine more factors like wind and snowfall into our clustering models.

1.1 Literature Review

In 2020, Bradatan et al. conducted research on the relationship between climate factors, including temperature and rainfall, and child health in Honduras. They used generalized estimating equations for binary logistic models and spatial association to analyze the child health data from the Honduras Demographic Health Survey 2011–2012 dataset and climate

data (1981–2012) from the Climate Research Unit (CRU TS3.21). Their results showed that areas experiencing significant temperature anomalies are also those with the worst child respiratory problems [3].

In 2021, Erick et al. conducted research on the impacts of rainfall variability and multiscale vulnerabilities on birth weight in Amazonian regions. Using Bayesian models, their results showed that rainfall variability confers intergenerational disadvantage, particularly affecting socially marginalized Amazonians in overlooked areas [4].

Occupational heat stress risk is influenced by various factors, including environmental conditions, heat sources, physical activity levels, clothing, and individual factors. Workload considerations, as outlined in the OSHA Technical Manual, play a crucial role. Prevention involves identifying heat hazards in the workplace and considering both environmental and metabolic heat sources. Employers should assess total heat stress, comparing it to published guidance and being mindful of heat advisories. Workers may experience heat stress at temperatures lower than public advisories, emphasizing the importance of thorough evaluation and proactive prevention measures. [5]

In the face of escalating extreme weather events linked to climate change, safeguarding construction workers becomes paramount for contractors aiming to ensure workplace safety and compliance with OSHA regulations. The challenges posed by extremely high temperatures include sunstroke, dehydration, and machinery-related risks, necessitating measures like facilitating access to water, scheduling shaded breaks, and promoting consistent use of personal protective equipment. Similarly, heavy rain introduces hazards such as slips, reduced visibility, and potential electrocution, demanding drainage planning, temporary work halts, and provision of appropriate protective gear. A well-structured construction schedule, coupled with preventive actions and worker education, emerges as a

crucial strategy to mitigate the adverse effects of weather conditions, ensuring a secure and productive construction site. [6]

Clustering algorithms are essential tools in unsupervised machine learning, offering insights into unlabeled data by identifying inherent patterns and groupings. Three main types of clustering algorithms—density-based, distribution-based, and centroid-based—provide diverse approaches to handling different datasets. Examples of popular clustering algorithms include K-means, DBSCAN, Gaussian Mixture Models, BIRCH, Affinity Propagation, Mean-Shift, OPTICS, and Agglomerative Hierarchy. These algorithms find applications in various fields, from fraud detection and customer segmentation to earthquake analysis and city planning. Selecting the appropriate clustering algorithm depends on the nature of the data and the specific goals of the analysis. The implementation of these algorithms, demonstrated using a sample dataset, showcases their distinct characteristics and use cases, emphasizing the importance of understanding their strengths and limitations in practical applications. [7]

K-Means clustering is a centroid-based algorithm designed to partition a dataset into K clusters, with each observation assigned to the cluster whose centroid (mean) is nearest. The iterative process begins by selecting the number of clusters (K) and initializing centroids, typically through random selection or more strategic methods. Data points are then assigned to the nearest cluster based on distance metrics like Euclidean Distance. The centroids are recalculated by averaging the data points within each cluster, and the assignment process is repeated until optimal centroids are achieved and assignments stabilize. This iterative nature parallels the Expectation-Maximization (EM) approach, involving steps of assigning data points to likely clusters (Expectation) and recomputing centroids (Maximization) using least square optimization. Centroid initialization methods, such as random selection and Naive Sharding, impact algorithm efficiency, with sharding offering improved execution time compared to random initialization. Overall, K-Means clustering proves effective in

identifying distinct groups within a dataset by minimizing the sum of distances between data points and their assigned cluster centroids. [8]

Clustering analysis, an unsupervised learning method, categorizes data points into distinct groups based on similarities, with various techniques such as K-Means, Affinity Propagation, Mean-shift, DBSCAN, Gaussian Mixtures, and Spectral clustering. Among these, Density-Based Spatial Clustering of Applications with Noise (DBSCAN) stands out for its ability to identify clusters with arbitrary shapes and handle noise and outliers effectively. DBSCAN relies on defining a neighborhood around each point and requires parameters like 'eps' for neighborhood distance and 'MinPts' for the minimum number of neighbors within the radius. It classifies points as the core, border, or noise based on their relationships. The algorithm proceeds by assigning clusters to core points and recursively expanding them to density-connected points. DBSCAN is particularly advantageous when dealing with non-spherical data or an unknown number of classes. In summary, DBSCAN's robustness to irregular shapes and noise makes it a preferred choice over traditional methods like K-Means in clustering analysis under diverse real-life data scenarios. [9]

The Silhouette Score is a crucial tool in evaluating the effectiveness of clustering algorithms by providing a quantitative measure of the cohesion and separation of data points within clusters. Calculated by assessing the average distances within and between clusters for each data point, the Silhouette Score aids in determining the appropriateness of cluster assignments. In Python, this metric can be easily computed using the silhouette_score function from scikit-learn. Interpreting the score involves considering its range from -1 to +1, where negative values suggest potential misclassifications, values close to 0 indicate ambiguous clustering and positive values signify well-defined and appropriately separated clusters. The Silhouette Score holds significance in assessing algorithm performance,

identifying anomalies, selecting optimal cluster numbers, and guiding clustering improvements. [10]

1.2 Workflow Diagram

1. Dataset Generation
    a. Filter child care centers by longitude and latitude
    b. Match temperature and rainfall data based on longitude and latitude
2. Statistics & Data Cleaning
    a. Exploratory Data Analysis
    b. Interquartile Range
    c. generate a cleaned training set
3. Clustering
    a. K-means
    b. Gaussian Mixture
    c. DBSCAN
    d. Spectral
4. Result Analysis
    a. Calculate each cluster's season's average rainfall and temperature
    b. generate rank based on the highest temperature in summer and the highest rainfall
5. Data Visualization

2 Data Description And Processing

2.1 Datasets

The goal of our project is to accurately identify childcare centers in Florida that face elevated risks of high temperatures and rainfall. Consequently, it is crucial to ensure the reliability of our datasets. In this regard, the dataset for childcare centers originated from the U.S. government, specifically the Department of Health & Human Services. Additionally, we sourced Land Surface Temperature (LST) data and rainfall data from NASA. For temperature, we utilized daytime LST data, and for rainfall, we employed Global Precipitation Measurement (GPM) data.

2.1.1 Childcare Center

The evolution of the modern workforce has seen a surge in parental involvement, necessitating a greater reliance on childcare services. This shift is attributed to the prevalence of dual-income households and the rise in single-parent households, fostering an amplified requirement for secure and dependable childcare alternatives. Concurrently, a burgeoning awareness has emerged regarding the pivotal role of early childhood education in shaping a child's developmental journey. Consequently, parents are actively pursuing high-quality childcare centers that provide safe havens and deliver enriching educational programs within stimulating environments for their children. As a result, the significance and reliance placed upon childcare centers have notably escalated in recent years, reflecting their pivotal role in supporting working families and early childhood development.

We chose to examine vulnerable childcare centers due to the impact of temperature and precipitation because of Florida's unique climatic conditions, characterized by fluctuating temperatures, periodic heavy rainfall, and the occasional occurrence of severe weather events

like hurricanes, creating a distinct environment for childcare centers. These weather variations pose substantial challenges to the operational resilience and safety of childcare facilities. Understanding the impact of temperature and precipitation on vulnerable childcare centers in Florida becomes imperative in ensuring their preparedness, adaptability, and ability to safeguard the children under their care. By analyzing these effects, we aim to uncover the vulnerabilities these centers face, ranging from infrastructure susceptibility to potential health risks for children. This investigation will not only elucidate the specific challenges but also pave the way for tailored strategies, proactive measures, and resilient infrastructure implementations. Ultimately, our analysis seeks to contribute insights that empower childcare centers in Florida to mitigate risks, enhance their resilience, and continue providing safe and reliable care amidst the diverse climatic conditions prevalent in the region.

The childcare data utilized in our analysis is sourced from the Homeland Infrastructure Foundation-Level Data (HIFLD). We obtained all child care centers data in Florida, 2023, and all datasets are formatted in CSV. Each dataset includes the following attributes for every child care center: OBJECTID, ID, NAME, ADDRESS, CITY, STATE, ZIP, TELEPHONE, TYPE, STATUS, POPULATION, COUNTY, COUNTYFIPS, COUNTRY, LATITUDE, LONGITUDE, NAICS_CODE, NAICS_DESC, SOURCE, SOURCEDATE, VAL_METHOD, VAL_DATE, WEBSITE, and ST_SUBTYPE. For our analysis, we exclusively require the geographical coordinates, namely LATITUDE and LONGITUDE, to support machine learning modeling and visualization purposes.

2.1.2 Temperature

The temperature data utilized in our analysis is sourced from the Land Surface Temperature Daytime datasets (LSTD) [11] provided by the NASA Earth Observation program (NEO) [12]. We gathered monthly average LSTD data, measured in Celsius degrees, spanning from

March 2013 to February 2023, and all datasets are formatted in CSV. Each data file is structured as a 2D matrix, where the rows and columns denote latitudes and longitudes, respectively. Each file corresponds to the global LSTD data for its respective month, featuring a resolution of 0.1 degrees. Consequently, the dimensions of the 2D matrix in each data file are 3600x1800, where each point in the matrix records the LSTD value corresponding to its (latitude, and longitude) coordinates. The separation between adjacent points in the matrix is 0.1 degrees.

To acquire the monthly average Land Surface Temperature Daytime (LSTD) data for each child care center, we match the (latitude, and longitude) coordinates of each center with the nearest coordinates recorded in the LSTD datasets. Subsequently, we utilize the corresponding recorded value as the temperature for the given childcare center.

2.1.3 Rainfall

The rainfall data utilized in our study also originates from the NASA Earth Observation program, specifically derived from NASA's Global Precipitation Measurement (GPM) datasets [13]. Similar to the Land Surface Temperature Daytime (LSTD) datasets, we collected monthly average rainfall data, measured in mm/month, spanning from March 2013 to February 2023, with all data files formatted in CSV. Each rainfall data file comprises a 2D matrix with dimensions 3600x1800, where the rows and columns correspond to latitudes and longitudes, respectively. Within this matrix, each point records the rainfall data associated with its respective coordinates. Maintaining a resolution of 0.1 degrees, each data file encapsulates the monthly average rainfall data for its corresponding month. Similar to our approach for temperature, when acquiring the monthly average rainfall data for each childcare center, we select the nearest recorded coordinate from the rainfall datasets as the value for the current center.

2.2 Data Processing

We utilized Python to link the geographical location data of childcare centers with NASA's rainfall and temperature data, creating a comprehensive dataset. Firstly, we generated a list of dates starting from March 1, 2013, to February of 2023, covering the first day of every month. Then, we read a CSV file containing the latitude and longitude information of childcare centers, and based on these coordinates, we calculated the corresponding row and column numbers on the global grid of NASA's rainfall and temperature data.

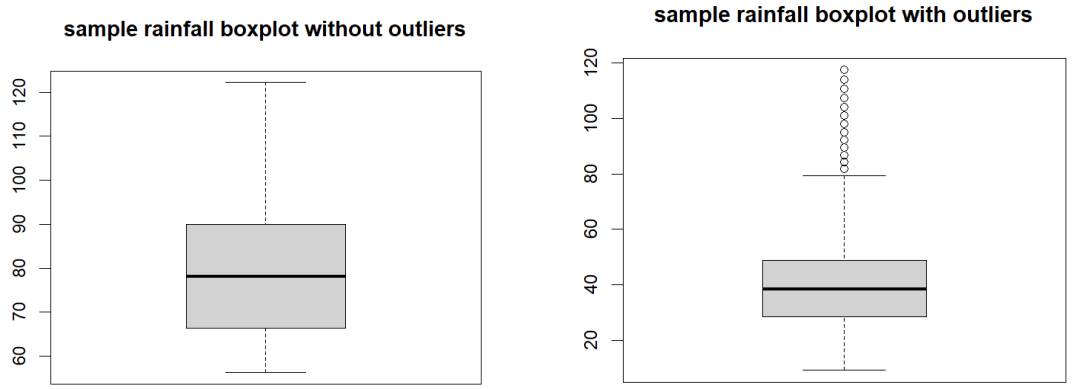
For each date in the list, we opened the respective NASA rainfall and temperature data file using Python. These files contained global rainfall and temperature data for the corresponding dates. We read this data and, using the previously calculated row and column numbers, extracted the specific data for each childcare center. We then combined this rainfall and temperature data with the ID of the childcare centers, forming a new data list. This list included the ID of each childcare center along with the rainfall and temperature data for the first day of every month from March 2013 to February 2023.

Finally, we wrote this list into a new CSV file. This file now contains comprehensive information on childcare centers combined with geographical location and time-series rainfall and temperature data, providing a valuable data resource for subsequent analysis. Through this process, we effectively merged two separate datasets, creating a new dataset that can be used for our further steps.

2.3 Data Cleaning

Data cleaning is an essential process to ensure the accuracy and reliability of our data analysis. Our project focused on integrating and analyzing data from the childcare care center, temperature, and rainfall datasets. We first conducted a thorough examination of all three datasets for NULL values and non-applicable data points. This step is to identify any

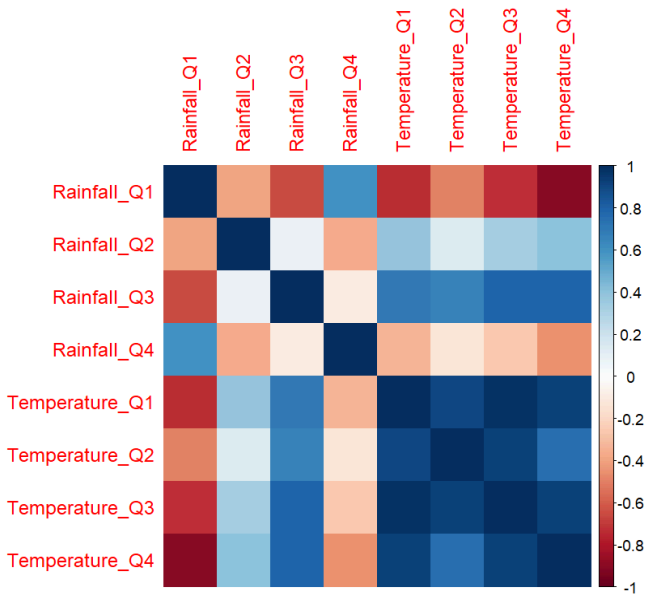
incomplete or irrelevant data that could skew our analysis. It was found that only the temperature dataset contains non-applicable data points with specific values like ‘99999.0’, these values represent the temperature in sea or undetected areas not relevant to our analysis. We removed all non-applicable values, this included all childcare centers with non-applicable temperature data. Using the remaining 4,307 childcare centers and their IDs as a reference, we merged the temperature and rainfall datasets for future analysis.



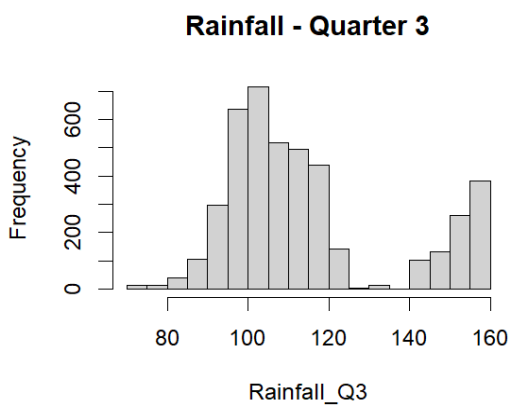
For each child care center’s 10-year temperature and rainfall data our team divided the data into four quarters, corresponding to the four seasons of the year. Seasonal segmentation was essential to identify seasonal patterns and abnormalities. We used the Interquartile Range (IQR) method to identify and remove outliers from each quarter of the temperature and rainfall data. Shown in the box and whisker plots above is a sample of the rainfall dataset before and after the removal of outliers. The first plot illustrates multiple circles or points falling outside the whiskers indicating outliers. These outliers are presented above the upper whisker indicating there are many instances where rainfall was significantly higher than the typical range. The second plot presents the data after outliers have been removed using the IQR method, as shown the resulting plot shows data points tightly grouped, and the whiskers are shorter reflecting a dataset that represents a more typical range of rainfall without extreme values. By continuously inspecting, cleaning, integrating, and refining the dataset we laid a solid foundation for our future data analytics to be accurate and insightful.

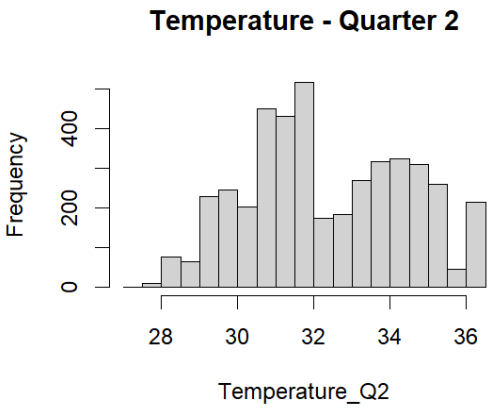
3 Exploratory Data Analysis

Exploratory data analysis is necessary for any data analytics project to understand the structure of the data, identify any patterns or abnormalities, and formulate hypotheses for further statistical testing. Our EDA consists of a variety of analytical techniques and visualization, including correlation matrix, histograms, kernel density estimates, and trend analysis over quarterly data.

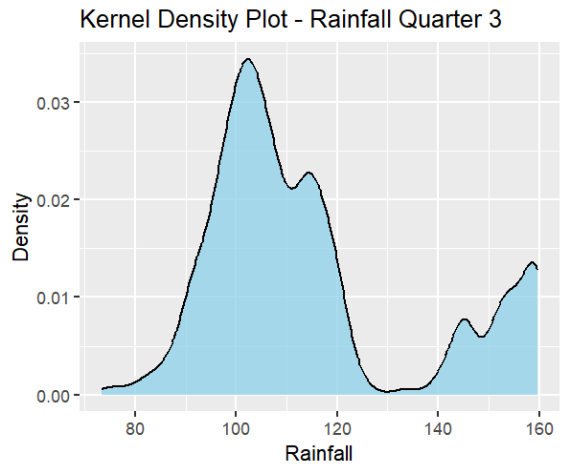


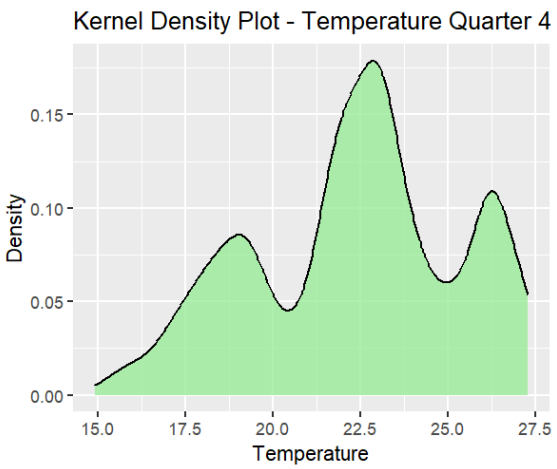
To begin our EDA process, a correlation matrix was developed to examine the relationships between the different variables within our dataset. The image shows a quantitative assessment of the strength and direction of linear relationships between pairs of variables. From the image, we can tell a strong positive correlation between the temperature in quarter 3 and quarter 4 represented by dark blue representing a positive correlation, and a strong negative correlation between rainfall in quarter 1 and temperature in quarter 4 represented by dark red.





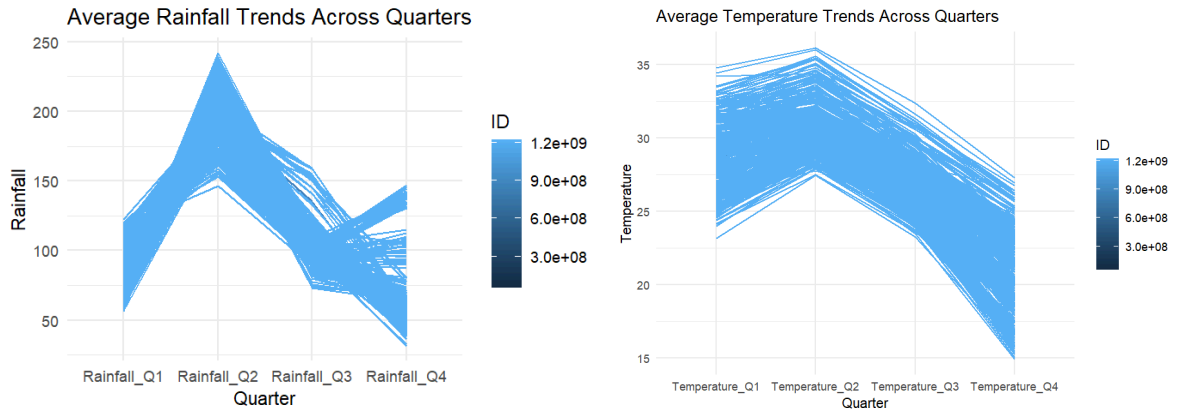
A series of histograms were generated to visualize the distribution of key variables. A sample of temperature and rainfall histograms are shown below. The histogram for Rainfall in Quarter 3 shows a normal distribution of rainfall with a slight skew toward the right indicating the concentration of data points around the median with some higher rainfall events less frequently occurring. The histogram for Temperature in Quarter 2 shows a multimodal distribution with several peaks meaning there may be multiple different common temperatures during this quarter.





Kernel density estimates (KDEs) were generated to obtain a smooth estimate of the data distribution. The Rainfall kernel density plot shows a large peak of around 100 and some minor peaks representing some common rainfall amounts; this may suggest typical weather

conditions in Quarter 3 or Autumn. The temperature kernel plot for Quarter 4 shows several common temperatures during this quarter indicating a transitional season with high variability in temperature.



Finally, we created two graphs to identify the trends across quarters, capturing variations and patterns. The graph depicts the variability and trends of rainfall across four quarters, each line represents a different location or ID and the spread of lines indicates variability in rainfall. For Rainfall, there is a noticeable peak in the second quarter which represents summer. The pattern suggests a cycle of rainfall with the second quarter corresponding to Florida's wet season. The peak in the second quarter or summer season for temperature and decrease in temperature in the third and fourth quarters are normal observations. The tightly grouped lines indicate the peak indicates the temperature is consistent across the childcare locations. Insights drawn from these graphs are valuable for further analysis. The Exploratory data analysis process helped visualize and analyze the data's characteristics and guide our analytics phase. Our team then moved on to model construction to further analyze the data at hand.

4 Model Construction And Results

In this study, we employed Machine Learning clustering models to categorize childcare centers in Florida based on their average temperature and rainfall from 2013 to 2022 for each quarter of the year. Subsequently, we identified childcare centers with a high risk of temperature and rainfall by examining each cluster's average, minimum, and maximum values. To enhance reliability, we utilized four distinct clustering models: Kmeans, Gaussian Mixture, DBSCAN, and Spectral Clustering. We then employed the silhouette score to assess the performance of each model and generated comprehensive statistics for their results, including the average, standard deviation, minimum, and maximum temperature and rainfall for each quarter of the year.

4.1 K-Means

K-Means [14] functions as a clustering algorithm, tasked with segregating a dataset into distinct clusters based on the inherent similarities among its data points. The initial phase involves the random selection of a predefined number of cluster centroids. Subsequently, each data point undergoes assignment to the cluster with the nearest centroid, determined by Euclidean distance. This step facilitates the grouping of data points based on their proximity to the current centroids. Following this assignment, the algorithm updates the centroids by computing the mean of all data points within each cluster. The iterative nature of this process continues until convergence, marked by the stabilization of both assignment and centroid values or the completion of a predetermined number of iterations. The primary objective of K-Means is to minimize the within-cluster sum of squares, ultimately forming compact, well-defined clusters. However, the algorithm's sensitivity to initial centroid placement necessitates multiple runs with varying initializations or the adoption of strategies like K-Means++ for a more intelligent start.

In essence, K-Means strives to uncover meaningful patterns within a dataset by iteratively refining cluster assignments and centroids. Despite its effectiveness, the algorithm's sensitivity underscores the importance of careful initialization strategies to ensure the attainment of robust and reliable clustering outcomes.

In this study, we utilized Kmeans to categorize childcare centers into four clusters. The silhouette score for the obtained result is 0.568, indicating an acceptable level of clustering quality. The table below illustrates the average temperature and rainfall for each cluster during every quarter of the year.

	Number of points	Temperature Average (Celsius)				Rainfall Average (mm/month)			
		Spring	Summer	Autumn	Winter	Spring	Summer	Autumn	Winter
cluster 1	1425	30.77	32.15	28.43	22.98	74.88	227.46	111.35	53.30
cluster 2	915	33.07	34.77	30.88	26.27	63.87	197.16	151.92	61.54
cluster 3	288	26.75	30.59	25.47	16.79	108.90	179.19	98.38	116.63
cluster 4	1679	28.86	31.48	26.88	20.64	85.56	191.11	100.01	58.01

Examination of the table above reveals that clusters 1 and 4 collectively account for the majority of childcare centers, totaling 3104 out of 4307 across all clusters. Notably, Cluster 2, comprising 915 centers, exhibits the highest average temperature throughout all four quarters, indicating an elevated risk of high temperatures for childcare centers in this group. On the other hand, Cluster 3, consisting of 288 centers, experiences the highest average rainfall during Spring and Winter. In Summer, Cluster 1 takes the lead in rainfall, while in Autumn,

Cluster 2 surpasses others in this regard. These findings provide insights into the childcare centers facing an elevated risk of rainfall during specific seasons.

4.2 Gaussian Mixture Model

Gaussian Mixture Models (GMMs) [15] operate by representing a dataset as a combination of multiple Gaussian distributions. Each Gaussian component in the mixture represents a distinct pattern or cluster within the data. The model assumes that the observed data is generated from a mixture of these Gaussian distributions, and it employs the Expectation-Maximization (EM) algorithm to iteratively estimate the parameters of these distributions. During the Expectation step, the algorithm assigns probabilities to each data point, indicating the likelihood of it belonging to each Gaussian component. In the Maximization step, the model updates the mean, covariance, and weight of each Gaussian based on these assigned probabilities. This process iterates until the model converges to a stable solution, effectively capturing the underlying structure and clusters within the dataset.

GMMs are versatile and find applications in various domains, including clustering, density estimation, and anomaly detection. They are particularly useful when dealing with complex datasets that exhibit multiple patterns or when traditional clustering methods like k-means are inadequate. GMMs provide a probabilistic framework, offering not only cluster assignments but also a measure of uncertainty, making them valuable tools for understanding the inherent complexity and structure of diverse datasets.

In our project, we once again utilize Gaussian Mixture to divide childcare centers into four clusters. The silhouette score of the results is approximately 0.562, which falls within an acceptable range. The table below provides the statistics for each cluster.

	Number of points	Temperature Average (Celsius)				Rainfall Average (mm/month)			
		Spring	Summer	Autumn	Winter	Spring	Summer	Autumn	Winter
cluster 1	292	26.73	30.56	25.45	16.80	108.76	179.03	98.26	116.13
cluster 2	1404	30.78	32.18	28.44	22.96	74.99	227.54	111.59	53.24
cluster 3	913	33.07	34.79	30.89	26.28	63.87	197.15	151.98	61.97
cluster 4	1698	28.88	31.47	26.89	20.70	85.28	191.56	100.00	57.93

The table reveals that the Gaussian Mixture produced clustering results similar to those of Kmeans. Clusters 2 and 4 contain the majority of points, totaling 3102 out of 4307 across all clusters. Cluster 3, with a size of 913, exhibits the highest temperature in all quarters. Regarding rainfall, Cluster 1, with a size of 292, has the highest values in Spring and Winter, while Cluster 2, with a size of 1404, records the highest values in Summer, and Cluster 3 registers the highest values in Autumn.

4.3 DBSCAN

DBSCAN [16], which stands for Density-Based Spatial Clustering of Applications with Noise, is a clustering algorithm that groups data points that are close to each other in a high-dimensional space. Unlike k-means, DBSCAN doesn't require a specific number of clusters beforehand and can identify clusters of arbitrary shapes.

The algorithm relies on two key parameters: Epsilon (ε) and MinPts. Epsilon represents the maximum distance between two samples for one to be considered in the neighborhood of the other, while MinPts is the minimum number of data points required to form a dense region.

These parameters play a crucial role in determining the characteristics of the clusters identified by DBSCAN.

DBSCAN categorizes data points into three main types: core points, border points, and noise points. Core points are those with at least MinPts neighbors within a distance of ε. Border points are within ε distance of a core point but lack sufficient neighbors to be considered core points themselves. Noise points, on the other hand, do not fall into any cluster and are treated as outliers. The algorithm proceeds through several steps, starting with the selection of an arbitrary unvisited data point. If the chosen point is a core point, a new cluster is formed, and all reachable points within ε distance are added to the cluster. If it's a border point, it joins an existing cluster, and if it's noise, it's marked as an outlier. This process repeats until all data points have been visited.

In this project, we randomly generated 20 epsilons in a range from 0.15 to 1.53 and iteratively generated minimum sample values from 2 to 20 for the DBSCAN algorithm. We used the itertools package to concatenate the epsilon values and minimum sample values into a full combination list. Next, we imported the StandardScaler class from the sklearn library to scale the training set to unit variance. By having the scaled training set, we loop through every parameter combination in the list we obtained above and put them into the DBSCAN model to get cluster results. Every time we trained the model with a parameter combination, we used the result and scaled training set to calculate the silhouette score and compared the score with other results to get the best cluster. In the end, the best score from DBSCAN is 0.35, which, although not as good as K-means, also reveals an acceptable result. The attached form lists the number of nodes in each cluster and each season’s average temperature and precipitation.

	Number of points	Temperature Average (Celsius)				Rainfall Average (mm/month)			
		Spring	Summer	Autumn	Winter	Spring	Summer	Autumn	Winter
cluster 1	893	33.16	34.88	30.97	26.33	63.95	197.42	152.12	61.72
cluster 2	3106	29.74	31.8	27.6	21.72	80.61	207.82	105.3	55.81
cluster 3	153	26.77	30.37	25.38	16.93	103.81	167.98	95	97.89
cluster 4	124	26.71	30.87	25.57	16.67	114.28	194.59	102.56	136.98
noise cluster	31	27.57	29.75	26.09	20.32	84.87	174.65	122.11	91.04

4.4 Spectral Clustering

Spectral clustering [17] functions as a dimensionality reduction and clustering technique, based on the eigenvalues of the similarity matrix of the data. The initial phase involves the construction of a graph representation of the data, where each node corresponds to a data point and each edge weight reflects the similarity between the nodes. Subsequently, the algorithm computes the Laplacian matrix of the graph and its eigenvectors. The eigenvectors corresponding to the smallest eigenvalues are used to project the data into a lower-dimensional space, where the clusters are more separable. Then, a standard clustering algorithm like k-means is applied to the projected data to obtain the final cluster assignments. The main objective of spectral clustering is to minimize the normalized cut, which measures the dissimilarity between the clusters and the connectivity within the clusters. However, the algorithm's computational complexity and parameter selection require careful consideration to ensure the efficiency and accuracy of the clustering results. Spectral clustering excels at

capturing complex cluster shapes and structures that conventional clustering algorithms might fail to detect. In essence, spectral clustering leverages the spectral properties of the data graph to perform clustering in a reduced space. Despite its limitations, the algorithm's flexibility and effectiveness demonstrate its potential for various applications such as image segmentation, hierarchical clustering, and vector embedding.

In this study, we utilized Spectral to categorize childcare centers into four clusters. The silhouette score for the obtained result is 0.389, indicating an acceptable level of clustering quality. The table below illustrates the average temperature and rainfall for each cluster during every quarter of the year.

	Number of points	Temperature Average (Celsius)				Rainfall Average (mm/month)			
		Spring	Summer	Autumn	Winter	Spring	Summer	Autumn	Winter
cluster 1	1026	27.52	30.08	25.92	19.86	86.18	191.79	100.47	58.06
cluster 2	899	33.14	34.86	30.93	26.31	63.96	197.40	151.99	61.67
cluster 3	458	27.61	31.50	26.11	17.11	102.46	179.07	98.57	95.14
cluster 4	1924	30.97	32.57	28.51	22.94	76.57	218.70	108.66	54.93

5 Conclusion And Data Visualization

5.1 Data Visualization

To visually evaluate geographical patterns of precipitation-based risk exposure across childcare facilities, we leveraged map visualization with markers indicating centers colored by assigned clusters. This enabled the graphical inspection of spatial correlations in the data. We could scan for regional concentrations of higher or lower vulnerability, pick out apparent outliers, and look for clusters forming spatial boundaries.

In addition, we plotted cluster precipitation averages over time and compared them as the rank across identified groupings. This facilitated the analysis of temporal dynamics and precipitation deltas between risk clusters. Complementing the geospatial views, these graphs revealed variations along seasonal and multi-year dimensions that were not otherwise visible.

Linking visual analytics directly to the clustered modeling outputs provided an accessible yet multidimensional perspective for interrogating the data. Both regional inspectability and temporal traces proved crucial for properly interpreting model-driven risk assessments in context. These data visualizations constituted the most powerful lens for actionable insights on differential climate impacts that could inform localized resilience planning.

1. Childcare Location:

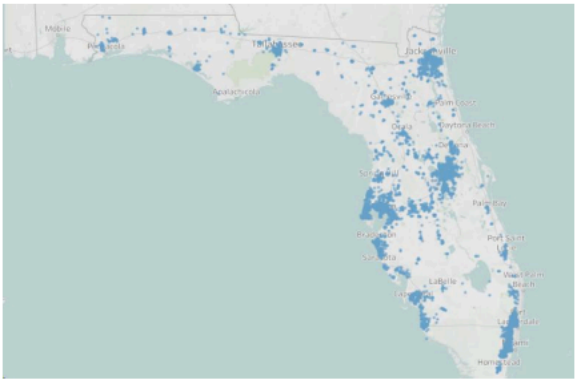


Fig 3. Childcare location clusters in Florida

2. K-Means Clustering:

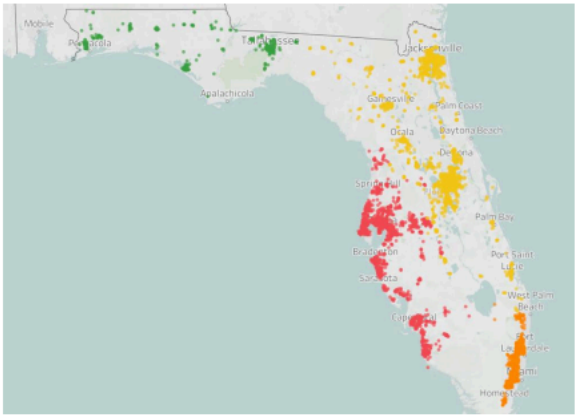


Fig 4. The clusters formed when visualized on a map using K-Means Clustering

3. Gaussian Mixture Model:

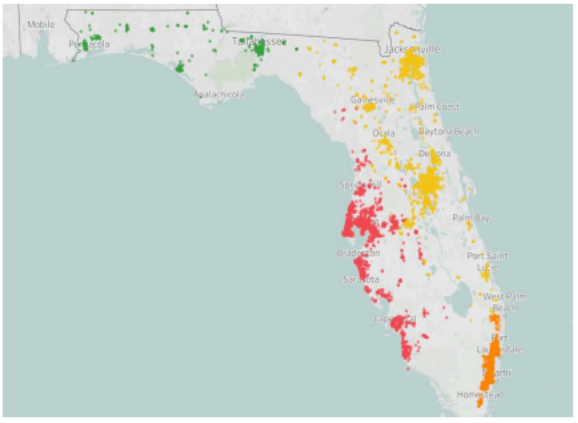


Fig 5. The clusters formed when visualized on a map using Gaussian Mixture Model

4. DBSCAN:

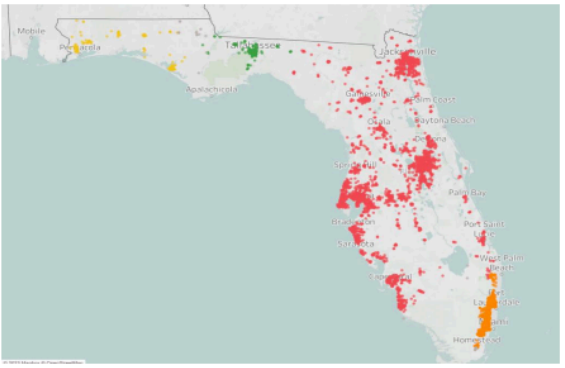


Fig 6. The clusters formed when visualized on a map using DBSCAN Clustering

5. Spectral Clustering:

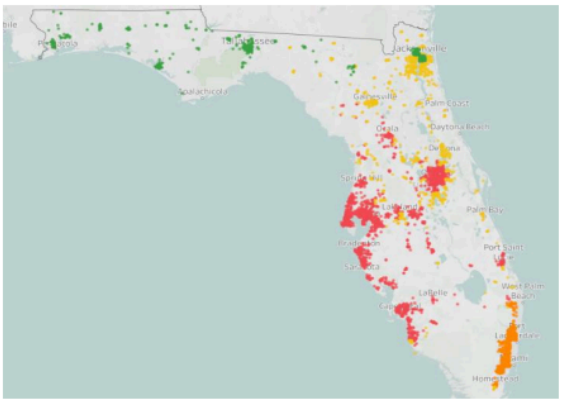


Fig 7. The clusters formed when visualized on a map using Spectral Clustering

6. Legend (Higher the rank value shows, the more increased the precipitation/temperature)

Cluster	Color	Rank
4	Red	4
3	Orange	3
2	Yellow	2
1	Green	1
-1	Grey	Noise

In the Data Visualization plots, we have four clusters on the map of Florida, and we use the maximum precipitation within one year as the rank to color the clusters. As the table described, we use red color to represent rank 4, orange color to represent rank 3, yellow color to represent rank 2, and green color to represent rank 1. The higher the rank value shows, the more precipitation in the region, which also means that childcare centers in that area are more vulnerable to climate effects from temperature and precipitation.

The clustered model outputs grouped over 4,000 childcare centers across Florida into four tiers categorized by precipitation-based climate vulnerability. Assigned risk levels derived from statistical similarity among facilities' historical precipitation exposures over 10 years. The highest risk tier (rank 4 coded red) emerged concentrated heavily in the southwest region, where childcare infrastructure contends with Florida's predominant storm tracks. With the ability to experience extreme tropical rainfall events, these communities face elevated climate threats from floods. Conversely, the northern interior part of the state around the capital clustered more commonly into lower precipitation/risk levels. However, modeled vulnerabilities should not be viewed as static. While interior locations see less extreme rainfall annually, they remain susceptible to severe storms and variability between years. These spatial patterns and risk differentiations can guide planning and interventions to harden childcare infrastructure against climate impacts proactively. From localized drainage improvements to larger grid resilience initiatives, clustering model outputs help inform where resources for climate adaptation may provide the greatest return on investment. Ultimately, clustering and visualization enabled data-driven assessment of precipitation-related vulnerabilities to support evidence-based policies that protect Florida’s children by making their care environments more disaster-resilient.

5.2 Conclusion

In this project, we successfully leveraged clustering algorithms to group childcare centers in Florida by their vulnerability to climate effects from temperature and precipitation. The clustering models generated coherent rankings of climate risk exposure across facilities. Visualizations of model outputs revealed clear geospatial patterns in the data.

The clustering methodologies provided actionable insights into the relative vulnerabilities of childcare centers at a point in time. However, enhancements could improve an understanding of evolving risks. Expanding the feature set with additional climate factors like soil moisture and population density may reveal deeper dynamics. Applying time series analysis to account for seasonal and multi-year trends could strengthen predictive capabilities.

While the current scope covers only Florida due to limited data access and research time, the intention is to scale nationally to provide decision-relevant information on climate threats to childcare infrastructure across the entire United States. With further development and support, this type of modeling approach could enable data-driven planning and resource allocation to strengthen community-level resilience. The ultimate goal is an early warning system that helps safeguard the well-being of children by mitigating climate hazards proactively.

REFERENCES

[1] ChildCare.gov. "Child Care Licensing and Regulations." ChildCare.gov, U.S. Department of Health & Human Services

[2] Child Care Technical Assistance. "Building and Physical Premises Safety in Child Care Settings." Administration for Children and Families, U.S. Department of Health & Human Services.

[3] Bradatan, Cristina, et al. "Child health, household environment, temperature, and rainfall anomalies in Honduras: a socio-climate data linked analysis." *Environmental Health* 19 (2020): 1-12.

[4] Chacón-Montalván, Erick A., et al. "Rainfall variability and adverse birth outcomes in Amazonia." *Nature Sustainability* 4.7 (2021): 583-594.

[5] Occupational Safety and Health Administration. "Heat Exposure: Hazards and Possible Solutions." OSHA, United States Department of Labor.

[6] GoCodes. "Weather Effects on Construction." GoCodes, GoCodes.

[7] Lin, Jovian. "8 Clustering Algorithms in Machine Learning that All Data Scientists Should Know." freeCodeCamp, 3 Mar. 2023.

[8] Natasha Sharma. "K-Means Clustering Explained." Neptune.ai, 8 August 2023

[9] Debomit Dey. "DBSCAN Clustering in ML - Density-Based Clustering." GeeksforGeeks, 23 May, 2023

[10] Samina. "The Silhouette Score: A Crucial Tool for Evaluating Clustering Algorithms." Educative.

[11] NASA Earth Observing System Data and Information System (EOSDIS). "MODIS/Terra Land Surface Temperature/Emissivity Monthly L3 Global 0.05Deg CMG V006." NASA, Goddard Space Flight Center, https://neo.gsfc.nasa.gov/view.php?datasetId=MOD_LSTD_M.

[12] NASA Earth Observing System Data and Information System (EOSDIS). "NASA Earth Observing System Data and Information System (EOSDIS)." NASA, Goddard Space Flight Center, https://neo.gsfc.nasa.gov/.

[13] NASA. "Global Precipitation Measurement (GPM)." NASA, https://gpm.nasa.gov/missions/GPM.

[14] Ahmed, Mohiuddin, Raihan Seraj, and Syed Mohammed Shamsul Islam. "The k-means algorithm: A comprehensive survey and performance evaluation." *Electronics* 9.8 (2020): 1295.

[15] Yang, Miin-Shen, Chien-Yo Lai, and Chih-Ying Lin. "A robust EM clustering algorithm for Gaussian mixture models." *Pattern Recognition* 45.11 (2012): 3950-3961.

[16] Schubert, Erich, et al. "DBSCAN revisited, revisited: why and how you should (still) use DBSCAN." *ACM Transactions on Database Systems (TODS)* 42.3 (2017): 1-21.

[17] Von Luxburg, Ulrike. "A tutorial on spectral clustering." *Statistics and Computing* 17 (2007): 395-416.