1 Introduction

1.1 Motivation

India, with its diverse geographic and climatic conditions, experiences varying degrees of water scarcity and quality issues across regions. Clustering analysis on a well water dataset allows for the identification of similar characteristics and trends among wells, aiding in the targeted allocation of resources and policy interventions. By categorizing wells based on water quality, quantity, and geological factors, authorities can tailor region-specific strategies for sustainable water management, groundwater recharge, and pollution control. The output of such clustering analyses provides a nuanced understanding of the spatial distribution of water-related challenges, empowering policymakers, environmentalists, and local communities to make informed decisions for ensuring reliable and clean water sources.

1.2 Problem Statement

The wellwater dataset contains valuable information about various attributes related to well water quality. The challenge is to analyze and understand the underlying patterns in the dataset through PCA and clustering techniques. Clustering aims to group similar observations together based on certain features and PCA is to gain a compressed representation of the data that captures the major patterns and variations, facilitating a more efficient analysis and interpretation.

The project aims to uncover meaningful patterns and extract essential information from the wellwater dataset, providing a more manageable and interpretable representation for further analysis and decision-making.

2 Description of the dataset and the analysis problem

The Indian dataset was taken from http://cgwb.gov.in/ground-water-quality ,and it contains measures of well water quality across 9076 areas of the country.In total 17 features were determined for each well .Areas covered almost major parts of the country.Address of each well water was also recorded for visualisation of the dataset.It hasn't been previously used in any online case study.However there are other related case studies that were refereed.

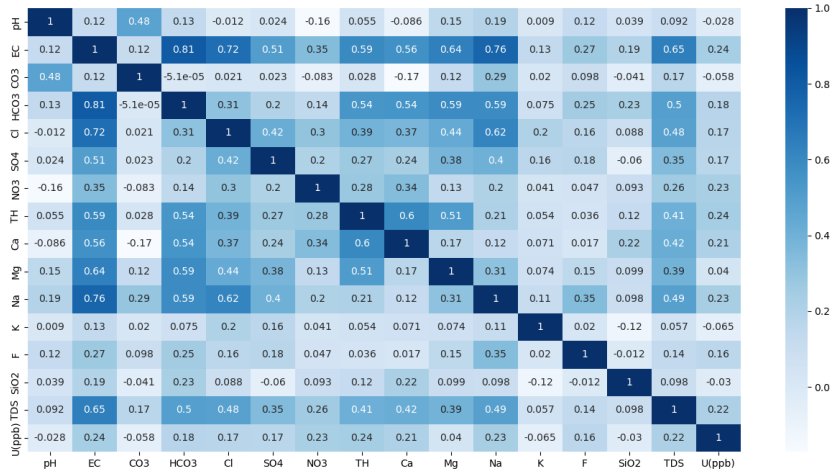
3 Analysis Description

3.1 Data Pre-processing

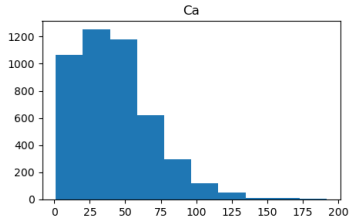
In the raw dataset,there were in total 15,647 cells which had null values.Some wells exhibited incomplete data or were found to be outliers and they were removed from the original dataset. Data that included values frequently lower than the detection limit of the method were excluded. When no recognition of ions was recorded, they were completed by the mean values of the neighboring data.Most of the features had different units of measure,so the range of values were differnet for each feature.So normalisation was done to ensure proper working of PCA . MinMax Scaler scaless the data within the given range, usually of 0 to 1,therefore it was used. To remove noise ,threshold values were given to each feature column,based on the domain knowledge.

3.2 Descriptive Analysis Of the Dataset

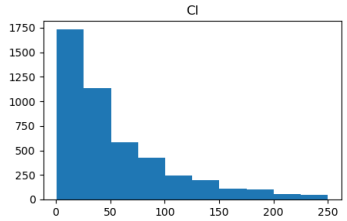
In the pursuit of ensuring access to safe and potable water, the Wellwater Dataset Project has endeavored to harness advanced data analytics techniques, namely clustering and Principal Component Analysis (PCA), to unveil patterns and insights within the diverse landscape of well water quality at various locations.



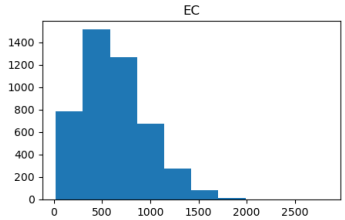
(a) Heat map of co-relation matrix



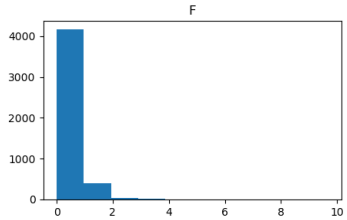
(b) Ca



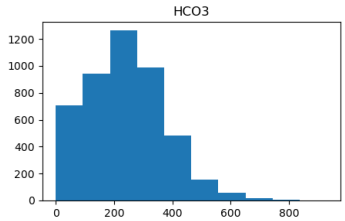
(c) Cl



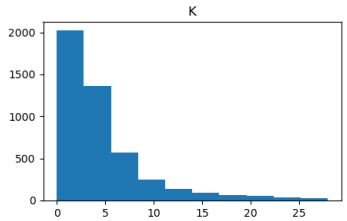
(d) EC



(e) Fluorine



(f) HCO3



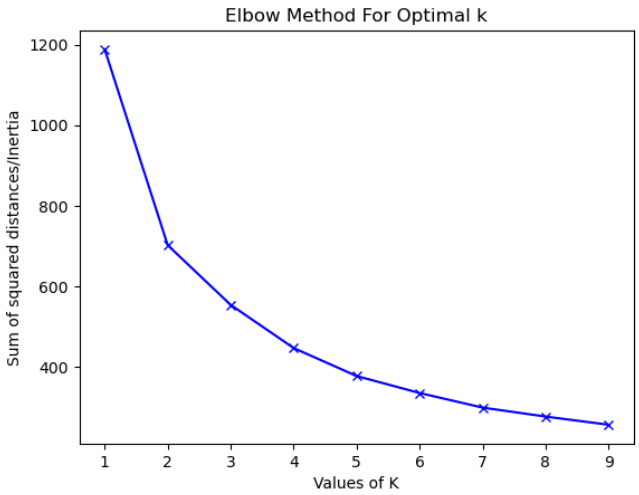
(g) K

3.3 Clustering Analysis

The application of clustering algorithms, such as k-means and fuzzy C-means clustering, has been instrumental in identifying natural groupings within the well water dataset. By grouping similar observations together, we have delineated distinct clusters that share common characteristics in terms of water quality attributes. These clusters not only highlight the heterogeneity in well water quality across different locations but also serve as a foundation for categorizing areas with similar contamination profiles. The interpretation of these clusters has unveiled valuable insights into the unique signatures of water quality issues.

Clusters identified through the clustering analysis pinpoint regions with similar water quality profiles, highlighting potential hotspots of contamination or areas requiring targeted intervention.

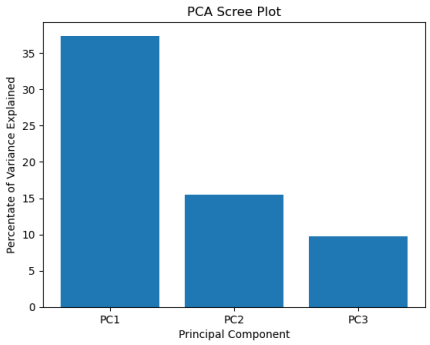
Through out the project,two clustering algorithms were applied to the data set,and clusters were obtained.To measure the effectiveness of each clustering techniques,silhouette score(a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation)) was determined.Clustering without doing PCA with K-means algorithm gave a silhouette score of 0.15,while clustering with PCA gave a score of 0.2951.Also Fuzzy C-means clustering with PCA applied,gave a score of 0.2873 .Therefore K-means with PCA was found to be best suitable.



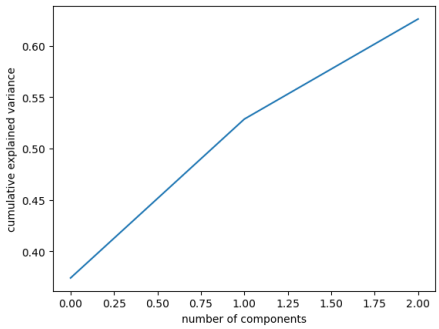
(a) Elbow method used for deciding optimum vale of K in K-means Clustering

3.4 Dimension reduction(PCA)

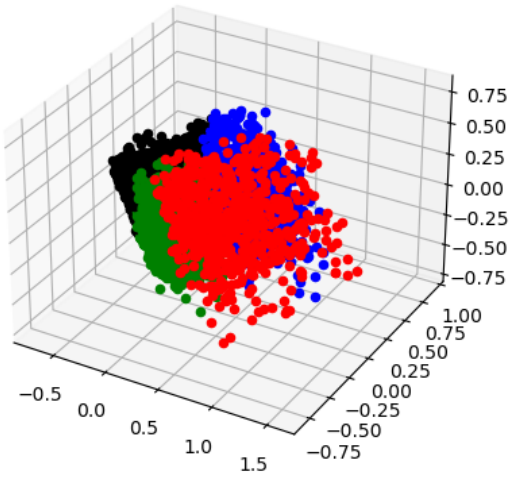
Principal Component Analysis has proven to be a powerful tool for reducing the dimensionality of the wellwater dataset. Through the extraction of principal components, we have created a condensed representation of the original data, capturing the essential patterns and relationships among variables. This reduced-dimensional space not only facilitates a more efficient analysis but also serves as a basis for identifying key contributors to water quality variations. The interpretation of principal components has illuminated the pivotal role played by specific variables in influencing water quality.The Principal component-1 was found to have higher influnce of NO3,Cl,TH features.Principal component-2 was found to have higher influence of TH,Na,and Cl.



(a) Scree Plot



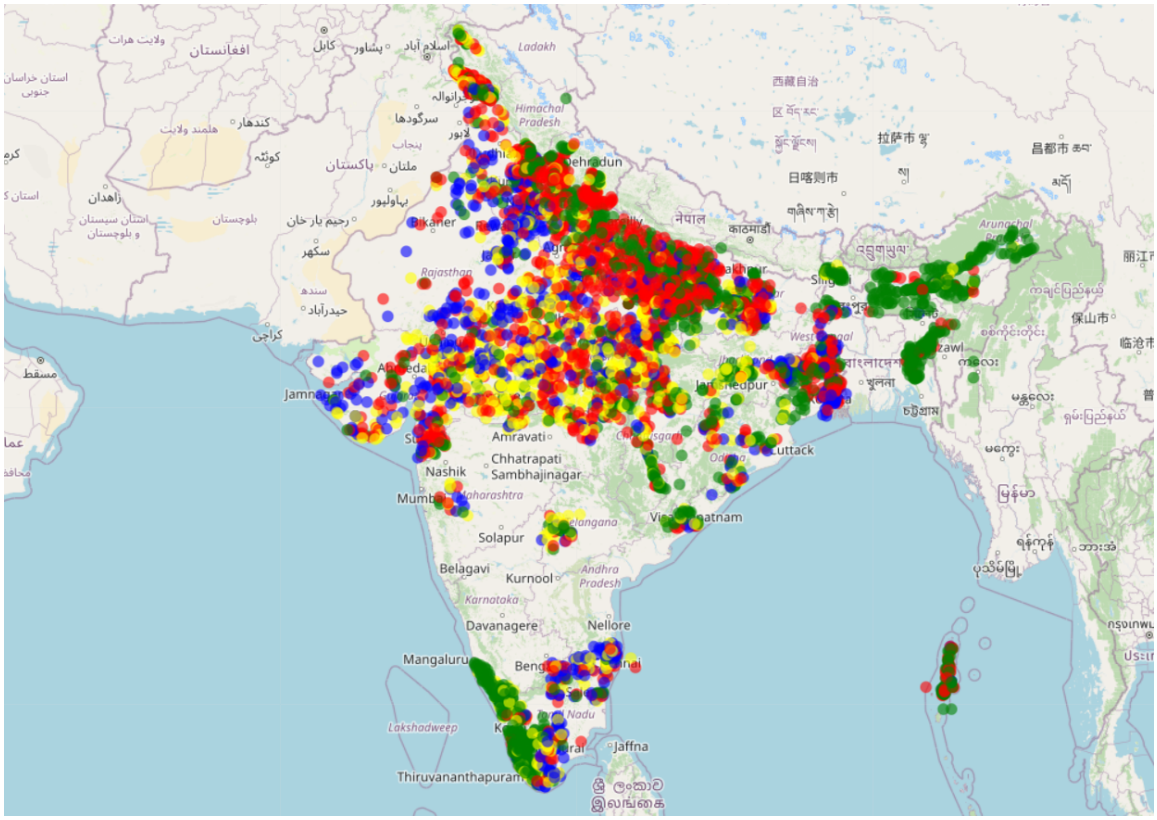
(b) Cumulative Plot



(c) Visualization of cluster in 3D after applying PCA

4 Results and Discussions

According to the results obtained ,in total 4 clusters were formed.Mean values of every feature of each cluster were noted,and comparing it with the domain knowledge,following results were determined.'PH' level in every cluster were distributed almost equally and hence didn't affect much.'EC' values of cluster-1 were comparatively much higher than the rest.Higher value of EC implies high amount of dissolved chemicals in water and hence treatment is required in those wells.Mean values of HCO3 in cluster-1 had a range of 370.89 ppm,which was higher than the rest clusters,which tells us that alkalinity level of wells of cluster-1 is more.Sodium(Na) and Calcium(Ca) values were also higher ,which makes these well water salty.Sulphate contents of cluster-1 wells were around 57 mg/L which can lead to a bitter taste, rotten egg smell, and cause diarrhea.Cluster-4 had NO3 values near 32 mg/L which is highly dangerous for feeding infants.It can also encourage the growth of algae and other organisms that give water a bad taste and odor.Cluster-2 wells had Total Hardness mean value of 74.4 mg/L which makes is considered as soft water.It had also a mean value of 12.3 mg/L for Magnesium which is good .Florine level of 0.6-0.7(found in cluster-2) has beneficial effects.Less SiO2 and Mg mean values is easy for treatment purpose.Total dissolved solids (TDS) values for cluster-2 were around 262 mg/L which is found to be excellent while cluster-3 and 4 values were acceptable,whereas cluster-1 had very poor values of 690 mg/L which can affect the taste and odor.Uranium values were found to be inside permissible range for cluster-2 (11.4 ppb),while for rest the values were concerning.Considering all the feature values of each cluster,it can be concluded that cluster-1 is a set of wells having more salinity,more hardness,and health-concerning quality,and hence should be governed carefully.Better policies should be implemented in these areas .



(a) Clustering without PCA,cluster-1 is red,cluster-2 is blue ,cluster-3 is green ,cluster-4 is yellow

5 Challenges

We tried different ways of scaling the features before feeding them to the classifier to train it better. Optimal sample size for training the model was a concern.Pre-processing was an issue due to missing of lattitude and longitude values of many locations,misplaced values of different feature values,and null cells in the dataset.Inferring conclusions from PCA was relatively difficult due to large number of data.

6 Analysis

This project have practical implications for policymakers, environmental agencies, and communities striving to ensure access to safe water. By understanding the unique challenges in different areas, targeted strategies can be

developed to address specific contaminants and improve overall water quality. The delineation of contamination zones serves as a roadmap for prioritizing resources and implementing measures to safeguard the well-being of communities.

7 Conclusion

The Wellwater Dataset Project has successfully leveraged clustering and PCA techniques to provide a nuanced understanding of well water quality across diverse locations. The identification of contamination zones based on these techniques offers a proactive and targeted approach to addressing water quality issues. As we move forward, these findings not only contribute to scientific understanding but also empower communities and stakeholders to make informed decisions for the sustainable management of well water resources.

8 Insight Learnings

The project underscored the importance of considering health implications in water quality analysis. Understanding the potential impact of variables such as pH, nitrate levels, and fluoride on health allows for targeted interventions to ensure access to safe and potable water. The wellwater dataset project provided valuable insights into the complexities of water quality, the application of clustering and PCA techniques.

9 References

- Indian well water(raw) dataset: http://cgwb.gov.in/ground-water-quality
- Bruce, Peter, Andrew Bruce, and Peter Gedeck. Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python. O'Reilly Media, 2020
- Fondriest Environmental, Inc. "Water Temperature." Fundamentals of Environmental Measurements. 7 Feb. 2014. Web. < https://www.fondriest.com/environmental-measurements/parameters/water-quality/water-temperature/ >
- Jolliffe, I. T. Principal component analysis. New York, NY: Springer, 2002.
- Introduction to Machine Learning with Python by Andreas C. Müller and Sarah Guido
- https://en.wikipedia.org/wiki/Principal$_c$*omponent*$_a$*nalysishttps* : *//stattrek.com/matrix−algebra/matrix−rank.aspx*