Final Report

Group 2

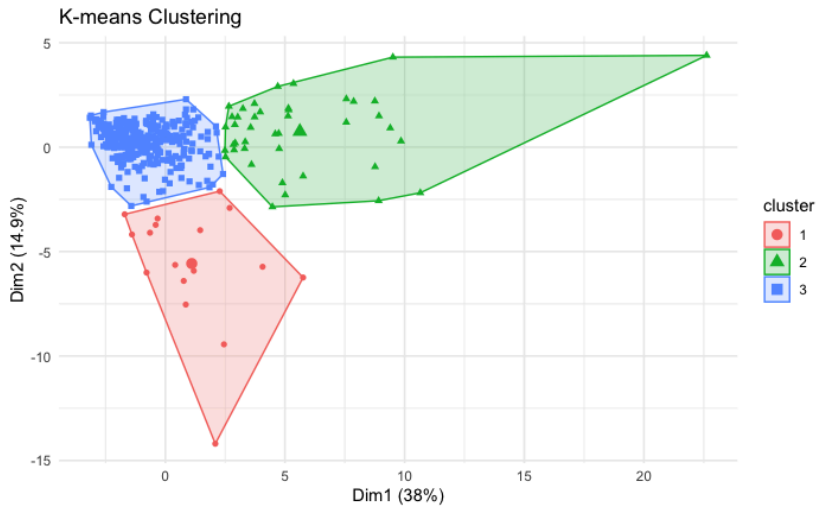
Zahia Khan
Arshdeep Singh
Muhammad Hammad

Table of Contents

Non-Technical Summary

The data analyzed in this study is a groundwater quality dataset that originated in 2020. The data was collected from districts in Telangana State, India. The dataset includes both categorical (non-numerical, grouped) variables and numerical variables. The categorical variables include District, Mandal, Village, and Season, which help distinguish the locations from which the data was collected. The dataset also classifies groundwater quality based on salinity and sodium levels across nine unique rankings, with C1-3 representing increasing salinity levels and S1-4 representing increasing sodium levels. Additionally, the dataset includes a classification for water suitability, indicating whether the water is P.S (Permissible), U.S (Unsuitable), or MR (Marginally Suitable) for agricultural and/or consumption purposes. The numerical variables in the dataset include latitude (Lat), longitude (Long), groundwater level (Gwl), pH, and a list of chemicals, minerals, and chemical combinations (SAR, TH, EC).

The groundwater data was grouped into three categories, or clusters, each representing different levels of water quality. These clusters were created based on the concentrations of various chemicals and minerals in the water. Cluster 1 had high levels of fluoride and sodium, which implied that the water quality in this group is poor and the water is likely contaminated by minerals. This water is not suitable for consumption without treatment. Cluster 2 showed high amounts of pollutants such as chloride, sulfate, and nitrates. Cluster 3 had low levels of contaminants, making it the best in terms of water quality and overall useability. This group is the most suitable for agricultural use and consumption without the need for additional treatment.



The K-means clustering graph shows the division of data into three distinguished clusters, all represented by a certain color and shape. The distinct separation helps visualize the varying levels of contamination in different groups. For example, one cluster might have higher levels of certain chemicals, indicating poorer water quality, while another cluster has lower contamination levels. Based on the graph and findings from the centroids, Cluster 3 (suitable water quality) has many centralized data points in comparison to the other two clusters that have widespread data points. It visually indicates that there is a lower contraction of substances such as chemicals or minerals that would contaminate the water. The k-means clustering graph helps to visually explain what areas have water that is safer for use. As a result of these findings, it is possible to conclude that some areas have water that might need treatment before use due to the presence of harmful chemicals, while the other areas have better quality water that are advised to be preserved and prioritized for consumption and agricultural use. This analysis in particular can help guide officials in Telangana State in their efforts to purify their water or conserve it.

Technical Summary

Exploratory Analysis

The exploratory data analysis (EDA) process focused on understanding the different parts of the groundwater quality dataset. The analysis involved the examination of both categorical and numerical variables. Initially, the EDA began with the creation of a correlation matrix to identify basic associations between the numerical variables. This step was essential whether it led to significant results or not, as it guided the direction of further analyses by demonstrating which relationships warranted further exploration, if any did. Originally, the analysis hypothesized that pH would be one of the more significant factors influencing groundwater level. However, after generating the correlation matrix, evidence supported the idea that pH had weak correlations with a majority of the other chemical elements, and that other factors were more likely to have a significant influence on groundwater level other than pH. Additionally, the elements that showed strong relationships with each other had already been grouped in meaningful combinations such as Total Hardness (TH) and Sodium Adsorption Ratio (SAR).

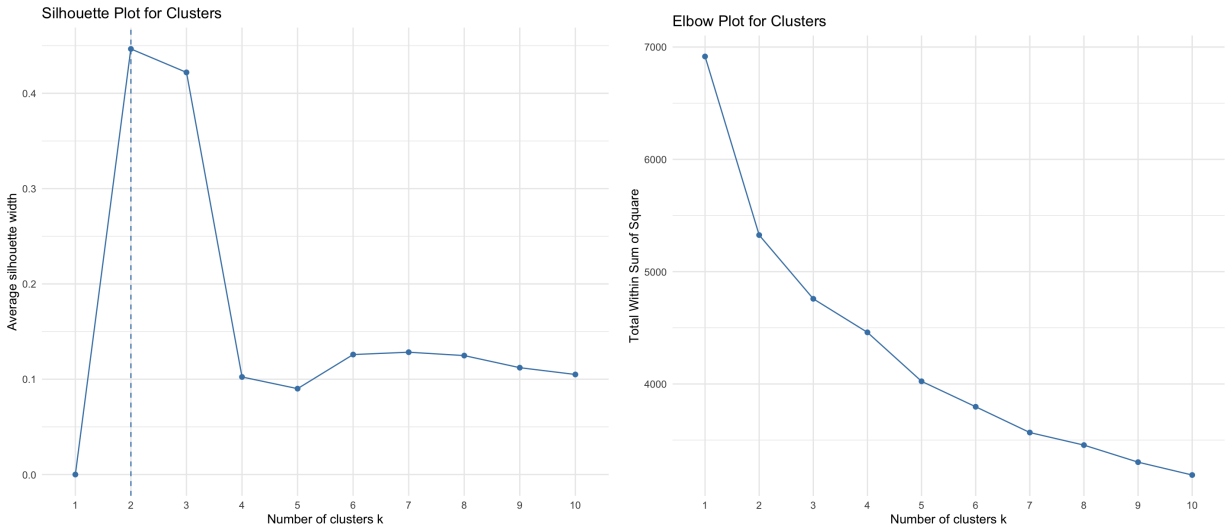
The most recurring element in the combinations, and the one with decent independent relations to groundwater level was sodium. To further investigate this relationship, a scatterplot between sodium and groundwater level was generated. However, this scatterplot did not provide sufficient insights. Following this, a Lasso regression model was conducted to see exactly how significant the predictors were to the groundwater level. The Lasso model concluded with a mean squared error (MSE) of 23.4899, which was nearly equal to the variance of groundwater level (23.52017). Even after applying a log transformation, the MSE increased incrementally to 24.976, proving that the results were still insignificant. After the significant amount of analysis done up to this point, the original goal of determining variations in groundwater level changed to a primary focus on identifying which factors cause the greatest variations in groundwater quality. There was a need to focus the analysis in the direction of groundwater quality, and to consider alternative methods of analysis that could incorporate categorical variables, as they had not been meaningfully investigated up to this point.

Given the ineffectiveness of the regression analysis, attention shifted towards examining the categorical variables. The categorical variable Season was removed first, as all entries were listed as “Post-monsoon 2020,” providing no meaningful information. Multiple Correspondence Analysis (MCA) was then performed on the categorical variables Mandal, Village, and District. The results showed that the first two dimensions, F1 and D2, explained only 0.54% and 0.59% of the variance respectively, totalling up to 1% of the variance. In a significant MCA, it is generally expected that the first few dimensions will explain at least 50% of the total variance. Even when the analysis was reduced to just the variable Village, the explained variance remained insignificant. The MCA led to the conclusion that the categorical variables had too much overlap and contributed little to no unique insights to the analysis. Overall, focusing on numerical predictors and Classifications is the most productive course of action in order to gain a proper understanding behind the variations in groundwater quality.

Application of First Line of Analysis

The first multivariate technique applied was the k-means clustering analysis. The initial step in this process was choosing the number of clusters. This was achieved by generating an elbow plot and a silhouette plot. The “elbow” of the elbow plot, which represents the point where adding more clustered would reduce the clustering quality, was found to be around k = 2 and k =

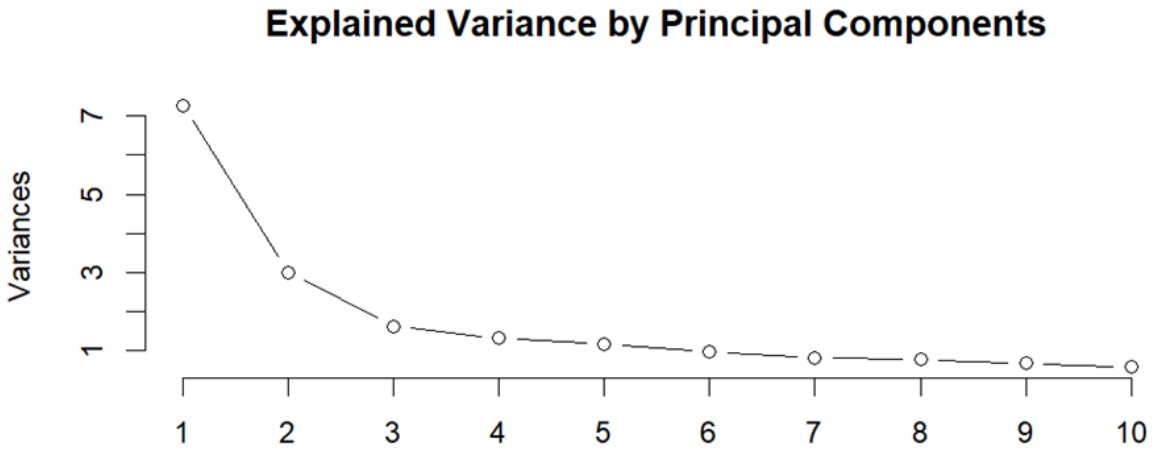
3, while the silhouette plot showed the highest silhouette width at k = 2. Ultimately, 3 clusters were chosen for the k-means clustering technique as a tradeoff, accepting a slight reduction in the simplicity of clusters for potential information that could be offered by an additional cluster.



The K-means clustering visualization (pictured in Non-Technical Summary) emphasizes the distinct boundaries between the clusters, showing that the groups have no overlaps between them. The axes, Dim1 and Dim2, represent principal components that account for 52.9% of the total variance. Each cluster corresponds to a different geographical region and groundwater quality level. With the values presented by the centroids, it is possible to determine the quantities of chemicals and minerals in each of the three clusters. Cluster 1 has the fewest data points, spans across several negative values, and has the characteristics of high levels of fluoride, sodium, TH, SAR, and bicarbonate. The presence of these minerals in high qualities implies that the water quality is unsuitable for consumption and agricultural use due to mineral contamination. Cluster 2, with the highest variability and values, shows elevated levels of electrical conductivity, TDS (total dissolved solids), chloride, nitrate, sulfate, calcium, and magnesium. The water quality is unsuitable in this case due to pollutant contamination. Cluster 3 is the most centralized and moderate group, with low levels of mineral and chemical content. This indicates that the water quality for areas that fall under Cluster 3 is suitable for both agriculture and consumption. Based on both the visualization and centroids, and applying the above information to a practical context, Cluster 1 requires treatment for usage, the areas in Cluster 2 would require significant changes in agricultural practices for suitable water, and Cluster 3 should be preserved and monitored for its high-quality water.

Application of Second Line of Analysis

The second line of analysis was PCA, the purpose of PCA was to reduce the dimensions of the data, as there were over 20 variables, it could be useful to get a more streamlined dataset.My initial PCA was at 10 PC . After realizing that 10 PCA wouldn't really help remove enough dimensionality, I decided to go with a middle ground, the elbow in the plot would show 2 Components, but it only ended up explain 59% of variance, I set my own parameter at 75% which lead to 5 components, which I think is enough of a cut for reduced dimensionality without losing too much information. This change was better as I got a better understanding of the data and the new PC variables. RSC has a large negative loading for PC1, but the strongest positive loadings for E.C., TDS, Cl, and T.H. With negative loadings, SAR, F, Na, HCO3, and RSC have the greatest influence in PC2,Ca and T.H have positive influences.PC3 is positively affected by pH, CO3, and long_gis, but negatively affected by gwl and HCO3,sno. PC4 is positively influenced by lat_gis and long_gis.PC5 has a negative effect from K and No3 and a positive one from co3. These values are the ones that are the most influential, but should still be changed based on the interactions of the variables for a better PC.




```
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6     PC7     PC8     PC9    PC10    PC11    PC12
Standard deviation     2.6951 1.7350 1.27465 1.14966 1.08868 0.98450 0.91098 0.87766 0.82440 0.7642 0.70633 0.6603
Proportion of Variance 0.3632 0.1505 0.08124 0.06609 0.05926 0.04846 0.04149 0.03851 0.03398 0.0292 0.02495 0.0218
Cumulative Proportion  0.3632 0.5137 0.59492 0.66100 0.72027 0.76873 0.81022 0.84874 0.88272 0.9119 0.93686 0.9587
                          PC13    PC14   PC15    PC16    PC17     PC18     PC19      PC20
Standard deviation     0.61497 0.59547 0.2897 0.07964 0.06074 0.006497 0.003107 1.334e-15
Proportion of Variance 0.01891 0.01773 0.0042 0.00032 0.00018 0.000000 0.000000 0.000e+00
Cumulative Proportion  0.97757 0.99530 0.9995 0.99981 1.00000 1.000000 1.000000 1.000e+00
```

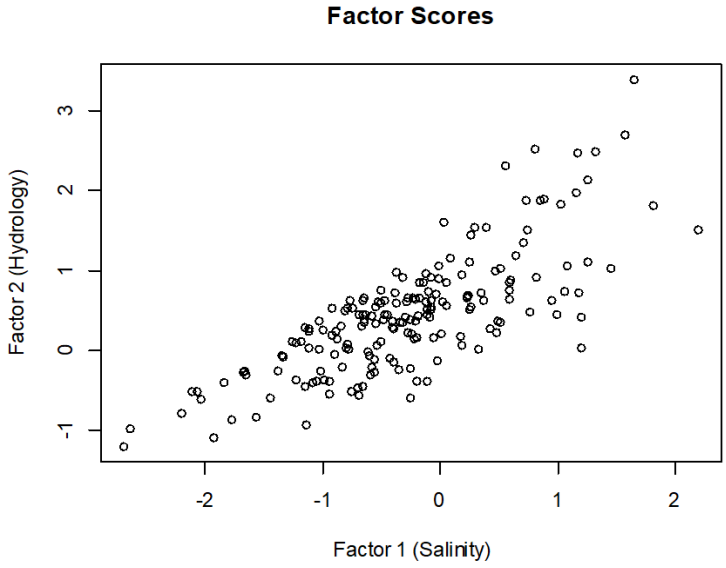
	PC1	PC2	PC3	PC4	PC5
sno	0.034726304	0.06004750	-0.149924044	-0.574044188	-0.12919377
lat_gis	-0.101693352	0.10018058	-0.141639145	0.518564027	0.11728153
long_gis	-0.049180789	-0.11688587	0.357116203	0.464803812	-0.05765706
gwl	-0.023147727	0.10596405	-0.343336295	-0.090026979	0.06040560
pH	-0.077443881	-0.29791328	0.479677383	-0.117792417	0.16711408
E.C	0.366429345	-0.06685667	0.001476911	0.023073545	0.02919335
TDS	0.366429345	-0.06685667	0.001476911	0.023073545	0.02919335
CO3	-0.057058490	-0.24800887	0.370043471	-0.349437520	0.24372463
HCO3	0.172348327	-0.33100797	-0.341764396	0.136890234	-0.06768074
Cl	0.347539630	0.03422401	0.070017554	-0.013951994	0.15117322
F	0.009619978	-0.38750090	-0.279270111	-0.026123118	-0.03230252
NO3	0.227759765	0.03611704	0.180288404	-0.014346728	-0.47445964
SO4	0.195026963	-0.04449464	-0.028705720	0.063689412	0.18333445
Na	0.279734012	-0.32691627	-0.045251950	0.032776522	0.03135094
K	0.104487556	-0.03136632	0.190301862	0.034922006	-0.72248585
Ca	0.276009712	0.22144844	-0.110080373	-0.066689196	-0.02155015
Mg	0.295417922	0.04983748	0.149689191	0.092796177	0.19191077
T.H	0.344238549	0.16383041	0.023261454	0.014841020	0.10198344
SAR	0.130828377	-0.49264572	-0.113379590	0.008332838	-0.01460535
RSC..meq....L	-0.281025608	-0.33040415	-0.160668065	0.029162785	-0.12437765

The first PC considered EC, and TDS variables, this led me to believe that this PC can be useful for understanding Salinity of the water and the minerals in the water. The next PC had the variables PH, F, SAR, this was useful for understanding the chemical makeup of the water, this would be useful for determining if the water was safe to drink. The combination of the first two PC would be ultimately useful because it could be used to determine if that water source is a viable source of water for agriculture and is safe for consumption. PC3 was made up for Carbonate, hand Potassium, and their interactions with each other is useful for determining water quality for maximizing crop growth. PC4 gave insight on how location effected water, this is ultimately useful for creating location based strategies for water treatment, the regions conditions are important to look at in this PC.PC5 was effected by Potassium, Nitrates, and Fluoride, high values in these variables can lead PC5 to shine light on the water quality and can be used as a measurement as the variables themselves in high concentrations have health risks, and PC5 can be used to identify water sources that need which treatment.

Application of Third Line of Analysis

The third line of analysis that was conducted was a factor analysis. The purpose of this factor analysis was to uncover the underlying structure of groundwater quality data collected post-monsoon in 2020. This dataset included critical hydrological and chemical properties such as groundwater level, pH, sodium, potassium, and sulfate levels. To prepare the data, missing values were handled by imputation, invalid entries were corrected, and skewed variables were

log-transformed to ensure suitability for factor analysis. Using the principal axis factoring method with Varimax rotation, we extracted two key latent factors. Adequacy tests supported this approach, with a KMO value of 0.72 and a highly significant Bartlett's test.



The results highlighted two dominant factors. Factor 1 represents chemical salinity and hardness, driven by variables like total hardness, sulfate, and sodium concentrations. Factor 2 reflects groundwater hydrology, characterized by groundwater level and pH variability. A scree plot confirmed the validity of these two factors, and visualizations like heatmaps and scatter plots demonstrated their interpretability. High Factor 1 scores indicated regions with higher salinity, while low Factor 2 scores were linked to deeper groundwater levels.

This analysis emphasizes the dual impact of chemical and hydrological influences on groundwater quality. Understanding these factors can guide targeted management strategies, such as mitigating salinity in high-risk areas or optimizing water resource availability. Overall, this factor analysis simplifies the complexity of groundwater data while preserving its essential insights, providing a strong foundation for decision-making and future studies.

Conclusion

The combined use of K-means clustering, PCA, and factor analysis has expanded the comprehension behind the variance in water quality and its impact on agricultural suitability. By using PCA values in tandem with the clusters found through K-means clustering, it is possible to

locate areas with optimal conditions for crop growth and minimal amounts of water contamination, as well as identify regions where water quality is deteriorating and develop targeted improvement strategies. Factor analysis further reduced groundwater data into two key factors, focusing on salinity, hardness, and hydrological variations. Together, these analyses can offer practical guidance, and its significance lies in its abilities to guide groundwater management strategies. Identifying the key factors that influence groundwater quality and categorizing them provides unique and valuable information for potential water treatment and conservation groups. Areas with specific chemicals or minerals contaminating the groundwater can be targeted for remediations, while regions with safe water can be managed to prevent future contamination or to supply consumable drinking water to areas that might have unsuitable water. These approaches can also be applied beyond Telangana, India, as long as similar trends are observed in the water quality of other regions. This research has the capability to support the actions and arguments made by environmental agencies, assist policymakers in creating regulations for agricultural practices, and help local communities as they take the next steps towards water sustainability efforts.

Appendix

I. **Group 2 Presentation -** https://depauledu-my.sharepoint.com/:p:/g/personal/zkhan59_depaul_edu/ER5GO96Ym_VCiiK8qNuLTyIB30ImaGl352xjuFXpBClLPw?e=tMTfhe

II. **Individual Report - Zahia Khan** https://docs.google.com/document/d/12AO6yCNUs3HOs5ePbRTlv5wRqmNi_-nB-2K3bC8LE5g/edit?usp=sharing

III. **Individual Report - Muhammad Hammad** https://depauledu-my.sharepoint.com/:w:/g/personal/zkhan59_depaul_edu/ERy0lFptcjxClwyaXIDNN6IB8vwTd6XXhBYXpx-YpTcW3A?e=2if8Z3

IV. **Individual Report - Arshdeep Singh** w4 (1).docx