Groundwater Analysis Presentation

Group 2
Zahia Khan
Arshdeep Singh
Muhammad Hammad

Data Overview

- 2020 groundwater quality dataset (Link)
- Collected from districts in Telangana State, India
- **Categorical**: District, Mandal, Village, Season
    - Groundwater quality classifications
        - Salinity/Sodium Levels: 9 unique rankings
          (C1-4 represent Salinity levels, S1-4 represent Sodium levels)
        - Water Suitability – P.S (Permissible), U.S (Unsuitable), MR (Marginal)
- **Numerical**: Lat, Long, Gwl, pH, various chemicals, minerals, chemical combinations

Cleaning & Transformations

- Removal of Season predictor
- Removal of any missing values
- Log transformations depending on the technique performed
- Employed scale() standardizations
    - Different variables with varying units (mg/L, meq/L, dS/m)
    - Predictors with longer ranges can dominate during calculations

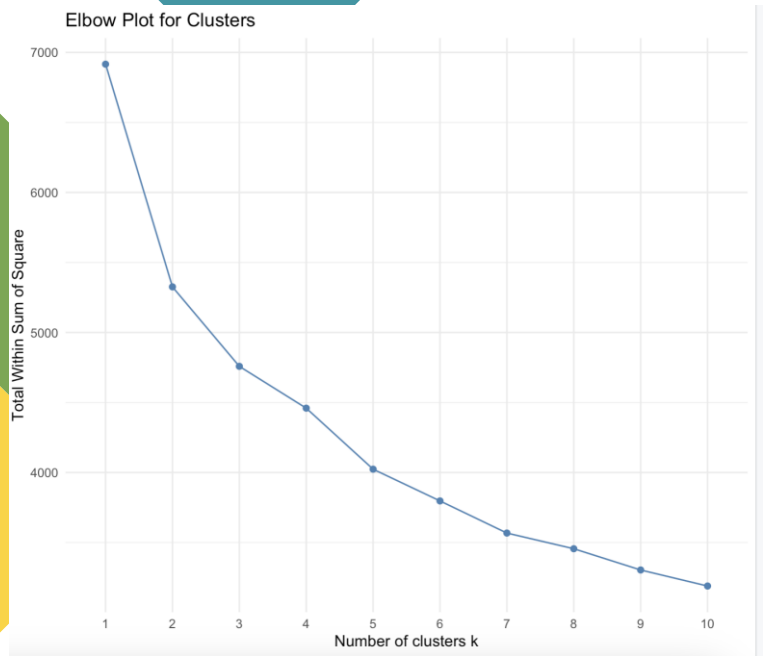
Multivariate Techniques

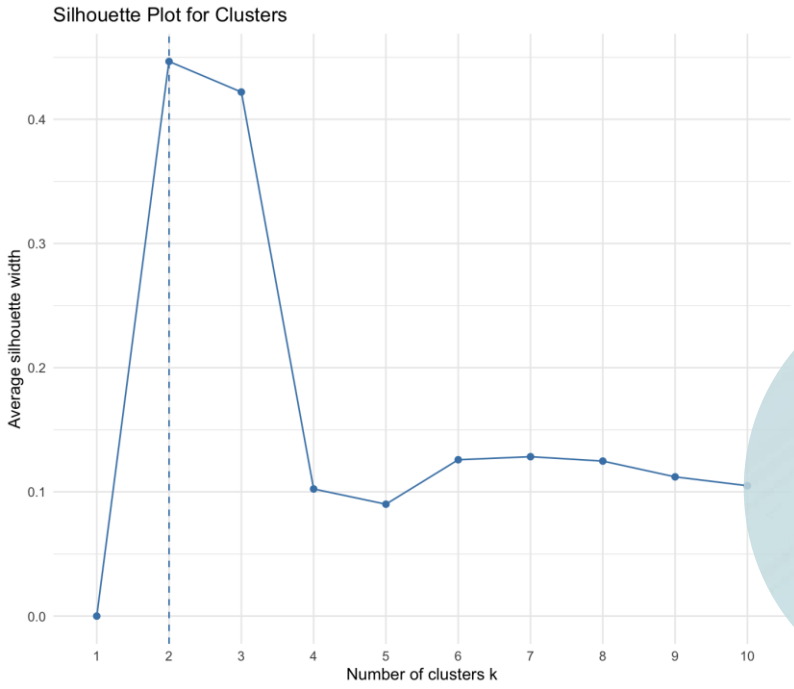
Goal: Understanding changes in groundwater quality by identifying trends in influencing factors

K-means Clusters

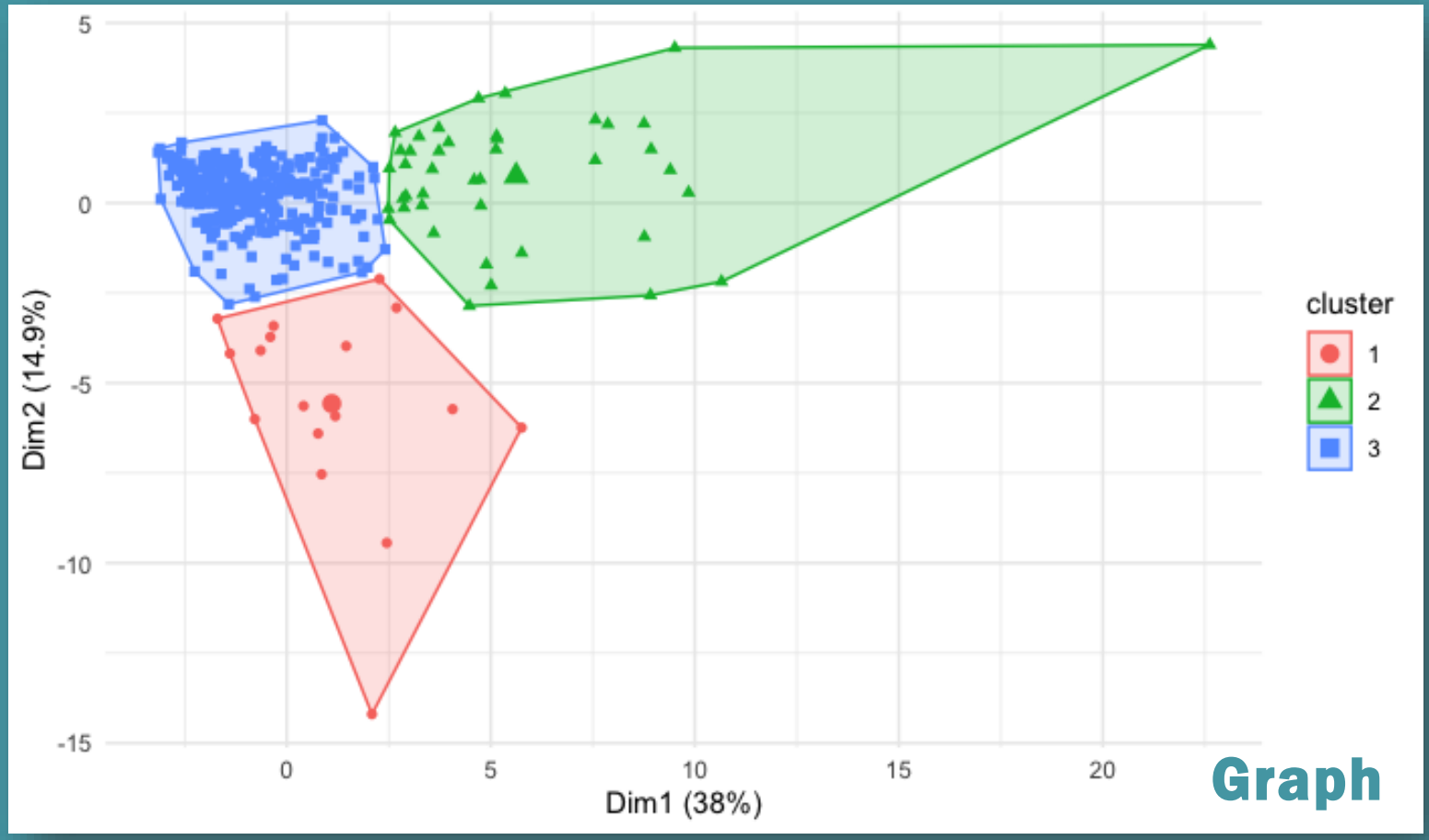
- Elbow Plot
  - Optimal point (elbow) at k = 2 to k = 3
- Silhouette Plot
  - Highest silhouette width is at k = 2
  - Falls off significantly after k = 3

- **Ultimately chose the cluster count of 3**
- Drop in silhouette plot width is not drastic
- Possible information and complexity can be gained from a third cluster
  - Tradeoff for a small reduction in overall cluster quality and simplicity





K-Means Clustering - Results



Graph

	sno	lat_gis	long_gis	gwl	E.C	TDS	CO3	HCO3	Cl	F	NO3	SO4	Na
1	-0.17827864	-0.6405536	0.28317723	0.0457561925	0.7336001	0.7336001	1.57132739	2.1699665	0.1705987	2.34203461	-0.1294726	0.2195428	2.1599723
2	0.28009808	-0.6352180	-0.28312710	0.0007569928	2.0132887	2.0132887	-0.27891789	0.5034775	1.9377518	-0.26254219	1.8033419	1.0800593	1.3479573
3	-0.02854047	0.1227731	0.02312851	-0.0026459117	-0.3170893	-0.3170893	-0.04901312	-0.1896584	-0.2754436	-0.09407783	-0.2403246	-0.1604402	-0.3050122

	K	Ca	Mg	T.H	SAR	RSC..meq....L
1	-0.04876560	-0.6944361	-0.2690816	-0.5814394	3.2486803	1.6444674
2	0.66442618	1.6042031	1.6981666	1.9897193	0.4301828	-1.8444339
3	-0.08848655	-0.1816050	-0.2181327	-0.2407965	-0.2395269	0.1617983

Centroids

K-means Clustering - Graph

- Post-scaling results
- Distinct boundaries w/no overlap
- Dim1 and Dim2 represent principal components that together explain about **52.9%** of the total variance
- Each cluster represents different geographical areas (location determined by Lat/Long Centroids)

- **Cluster 1:**
  - Spread across negative values
  - Least amount of data points
  - Most unique characteristics
- **Cluster 2:**
  - Most variability and high Dim1 values
- **Cluster 3:**
  - Centralized and moderate

K-means Clustering - Centroids

(↑ LVL) **Cluster 1:**

- Fluoride and Sodium
- TH and SAR
- Bicarbonate

(↑ LVL) **Cluster 2:**

- Electrical Conductivity, TDS, and Chloride
- Nitrate and Sulfate
- Calcium and Magnesium

(↓ LVL) **Cluster 3:**

- Mineral Content and Salinity
- Calcium and Magnesium
- Fluoride and Sodium

Concluding:

- ✓ Unsuitable water quality
- ✓ Mineral contaminants
- ✓ Can be treated for use

Concluding:

- ✓ Unsuitable water quality
- ✓ Pollutant contaminants
- ✓ Requires scaled changes

Concluding:

- ✓ Suitable water quality
- ✓ Low levels of contaminants
- ✓ Consumable and useable

Principal Component Analysis

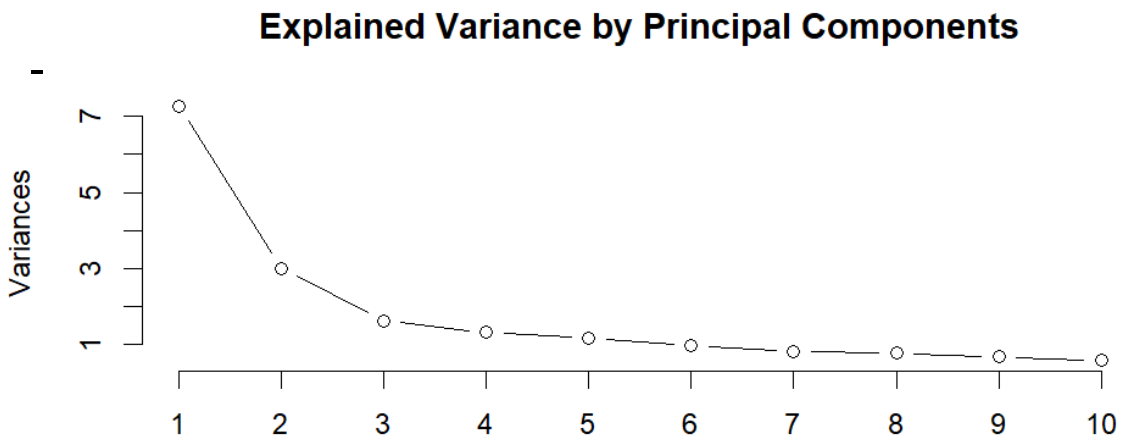
-Goal: Reduce dimensionality in our data set to better understand our variables and their interactions and relationships.

The Categorical Variables were removed for this PCA because they did not seem important as they were location based.

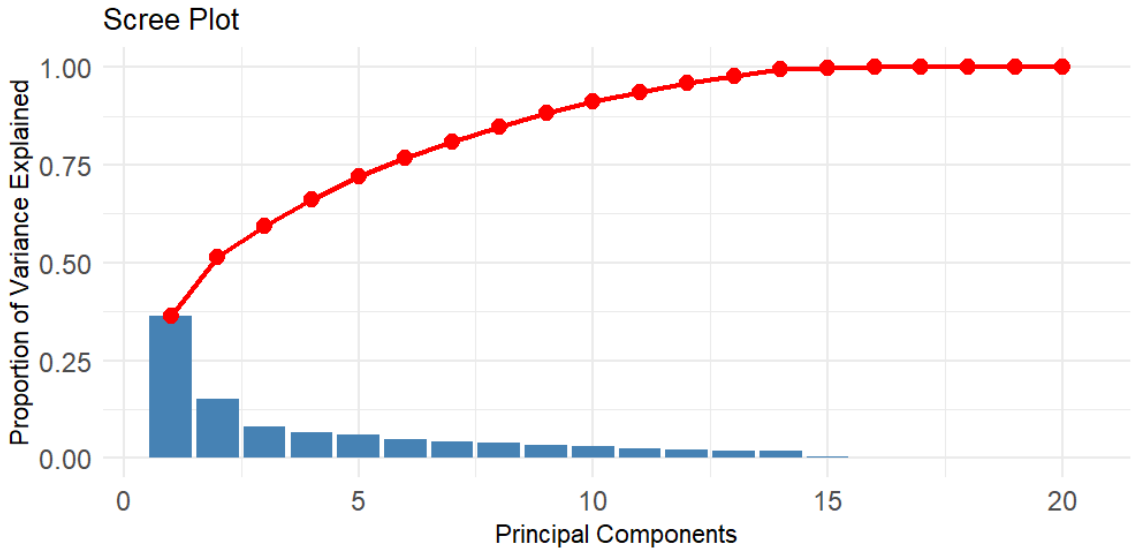
Overview of data:

The dataset has the following numerical variables: lat_gis (Latitude), long_gis (Longitude), gwl (Ground Water Level), pH (pH Level), E.C (Electrical Conductivity), TDS (Total Dissolved Solids), CO3 (Carbonate), HCO3 (Bicarbonate), Cl (Chloride), F (Fluoride), NO3 (Nitrate), SO4 (Sulfate), Na (Sodium), K (Potassium), Ca (Calcium), Mg (Magnesium), T.H (Total Hardness), SAR (Sodium Adsorption Ratio), and RSC..meq....L (Residual Sodium Carbonate).

Plots



Initial Plot shows 2 or 3 Components are to be used, Elbow is at 3.



The new plot with Variance shows around 5 are needed for about 75% variance explained.

Plots explained

PCA 1	PCA 2	PCA 3	PCA 4	PCA 5
Strongest Influence Variables :	Strongest Influence Variables:	Strongest Influence Variables:	Strongest Influence Variables:	Strongest Influence Variables:
EC,TDS	PH,F SAR	C03,Ph,k	Lat,Long,C03,	K,N03,F

Positive Influence:

Negative Influence:

	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

Overall, PCA1 and PCA 2 explain a majority of the variance in the dataset, with variables like EC,TDS,PH,F,SAR, at 59% of variance. The next three PCA scrape up more variance and take the total to about 75% variance, these last three PCA catch the smaller relationships in the data set.

Conclusion - PCA

- PCA1 and PCA 2 values in tandem can help understand water quality and be useful for determining whether or not water should be used for agriculture .
- PCA 3 can be used to manage water hardness and alkalinity, can be useful for best crop growth potential.
- PCA 4 Can be used to determine areas where water quality suffers and needs to be improved and be used to determine different strategies to improve water in different areas.
- PCA 5 can be used as a way to eliminate potential bad water sources to be used for agriculture.

Factor Analysis – Objectives and Methodology

- Purpose: Explore underlying groundwater quality structure and reduce dimensionality.
- Preprocessing: Handled missing data, corrected pH values, and normalized skewed variables.
- Method: Principal Axis Factoring (PAF) with Varimax rotation.

Adequacy tests:

- Kaiser-Meyer-Olkin (KMO): 0.72 (acceptable).
- Bartlett's Test: Significant ($p < 0.001$).

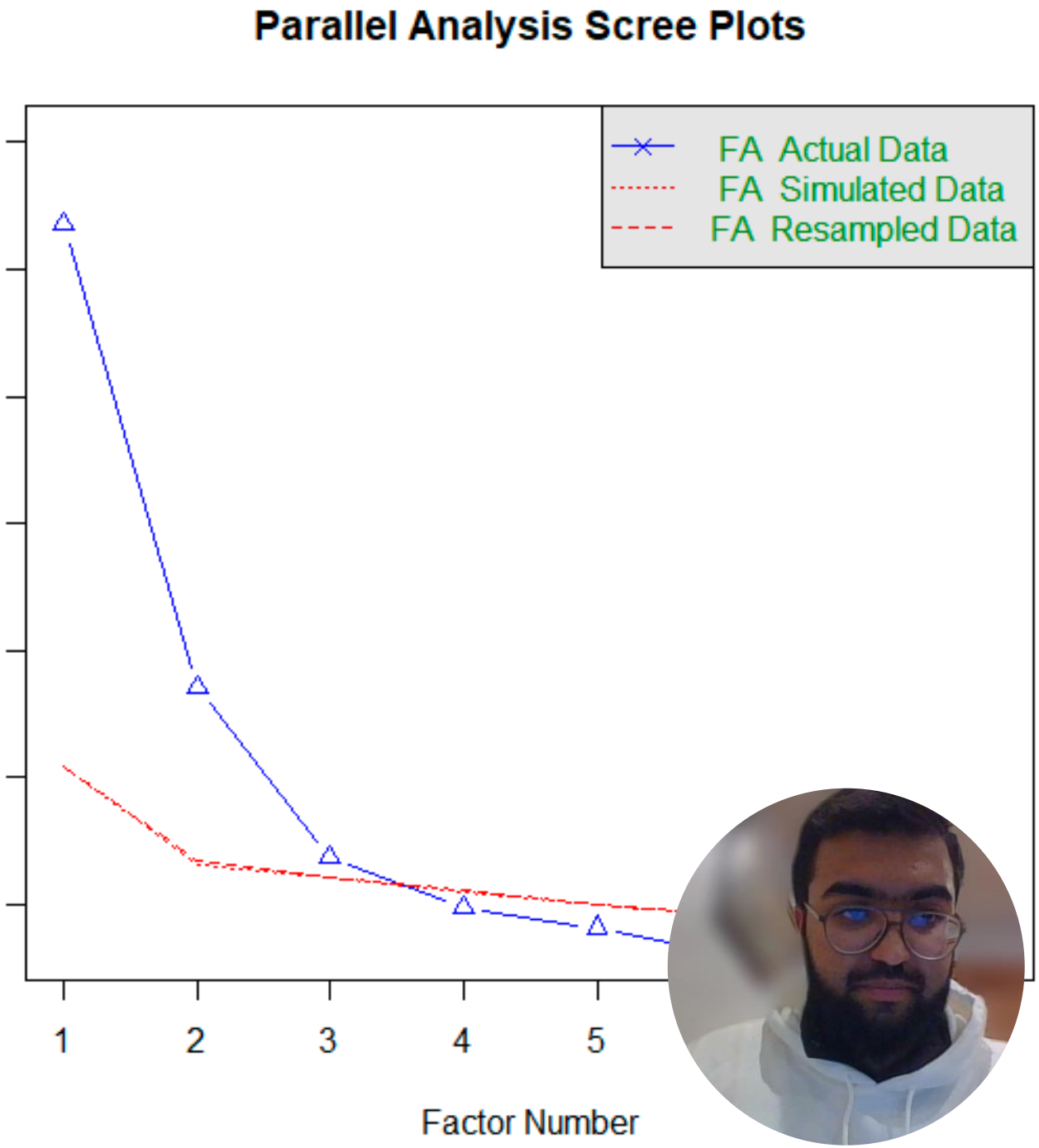
Factor Analysis – Results (Factor Loadings)

- **Factor 1**: Chemical salinity and hardness.
- High loadings: T.H (0.85), log_SO4 (0.76), log_Na (0.72).
- **Factor 2**: Groundwater hydrology.
- High loadings: GWL (-0.78), pH (0.56).
- Moderate contributions: log_K and log_RSC, indicating mixed influences.

Factor Analysis - Visualizations

Scree Plot:

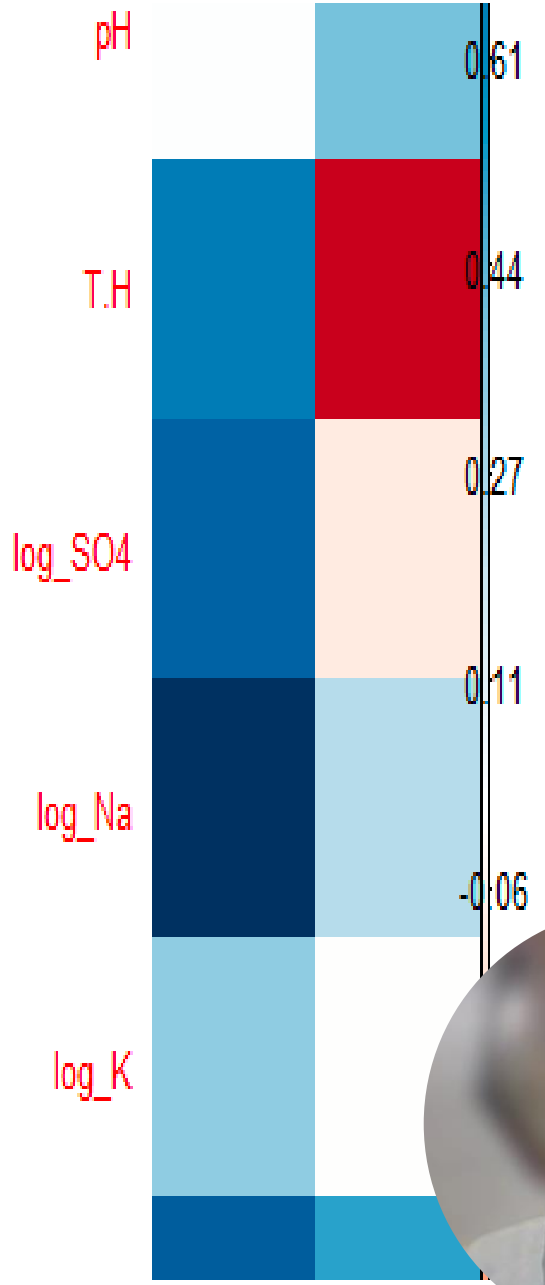
- Elbow at Factor 2, confirming two factors.

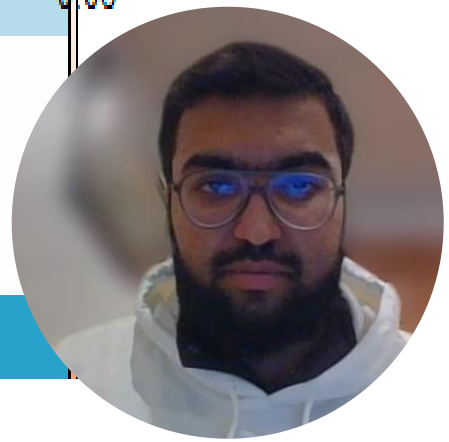


Factor Analysis - Visualizations

Heatmap of Factor Loadings:

- Visual grouping of variables by factor.

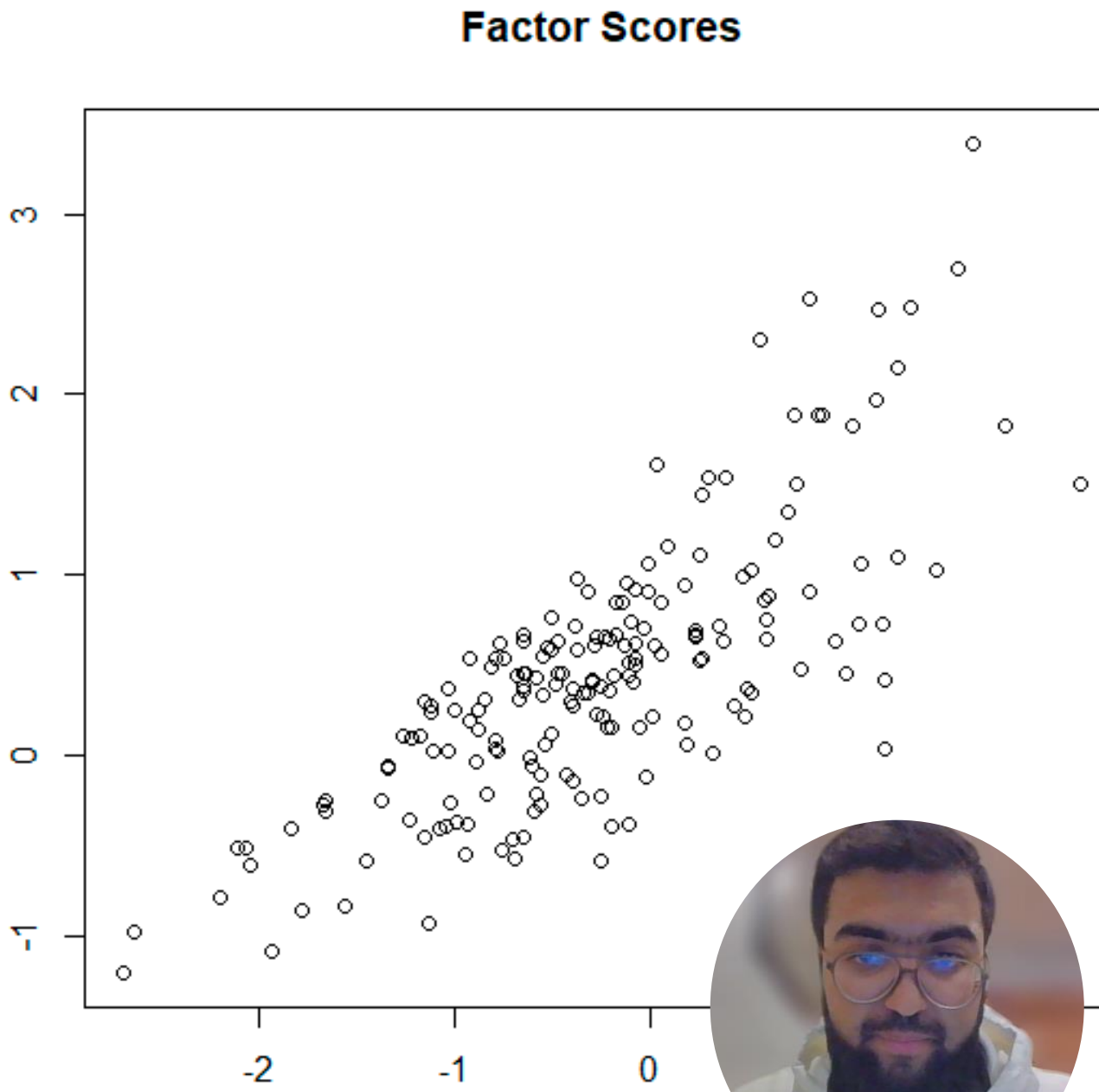




Factor Analysis - Visualizations

Scatter Plot of Factor Scores:

- High Factor 1: High salinity and hardness.
- Low Factor 2: Deeper groundwater levels.



Factor Analysis – Discussion

- **Factor 1 (Salinity and Hardness):**
- Dominated by dissolved ions.
- Reflects geological and anthropogenic influences.
- **Factor 2 (Hydrology):**
- Represents groundwater availability and pH balance.
- Highlights hydrological variations across regions.

Applications:

- Targeted water management strategies.
- Identification of regions needing salinity interventions.

Factor Analysis – Conclusion

- Factor analysis simplified complex groundwater quality data into two interpretable dimensions.
- Emphasized the dual role of chemical and hydrological factors.
- Facilitates decision-making for resource management and future research directions.

Overall Conclusions

- Combined technique usage allows for a better understanding of factors influencing water quality and impact on local communities
- PCA and K-means clustering identify areas with optimal water conditions and regions needing targeted quality improvements
- Factor analysis identifies variations in salinity, hardness, and hydrological variations
- Results can potentially be applied independent of region if conditions are met
- Analyses can guide groundwater management strategies, and support agencies and policymakers in water sustainability efforts