Arshdeep Singh

Individual summary

EDA.

The first thing that I did was clean the data for bad or missing values, from analyzing the raw data, the one thing that I found was that the PH had one bad number, it was coded as 8.. instead of 8., which meant that the data variable was being shown as a CHAR instead of a numeric value, I changed this value and then converted PH into a numeric value. Another I did for my data run was delete the region variables. The reason that I removed these values was because they were not providing enough information to the model, I did keep one of the region variables to get some insight, and I kept the Coordination variables as well. Another thing I noticed was 3 missing values in the GL variable, for this I just removed the three rows that were missing Running a histogram highlighted the distribution of the variables, from this I realized that the variables that needed to be transformed were the LONG and LANG which are coordination, which means they should not be changed as changing them can lead to misleading results. Another thing that I did was remove the X variable as all the values read NA which means it was a constant.

2.Inital PCA Model.

The first thing that I did after clearing up the data was standardize the data. Then I tried to run the PCA but kept running into some issues with the data points, it seemed like a few points were missing, so I used a mean average for the points that were missing. I removed the categorical variables as the two that were in the model were about the quality, and I wanted to really focus on the actual parts of the water and the specific of the water. After running my PCA, I set my cutoff as 90% of variance explained, which leads to 10 Principal components needed. 10 Components means 91% of variance explained. The variables strengths are shown below. The variable sno has a loading of 0.111 on PC10, showing a weak positive influence on this principal component. The variable Lat gis has a loading of -0.067 on PC10, showing a weak negative influence on this principal component. The variable Long Gis has a loading of -0.074 on PC10, showing a weak negative influence on this principal component. The variable gwl has a loading of 0.106 on PC10, showing a weak positive effect on this principal component. The variable ph has a loading of -0.104 on PC10, showing a weak negative influence on this principal component. The variable E.C has a loading of 0.053 on PC10, showing a very weak positive influence on this principal component. The variable TDS has a loading of 0.053 on PC10, showing a very weak positive influence on this principal component. The variable C03 has a loading of -0.171 on PC10, showing a moderate negative influence on this principal component. The variable HCO3 has a loading of 0.025 on PC10, showing a very weak positive influence on this principal component. The variable C1 has a loading of 0.106 on PC10, showing a weak positive influence on this principal component. The variable F has a loading of -0.730 on PC10, showing a strong negative influence on this principal component. The variable NO3 has a loading of -0.028 on PC10, showing a very weak negative influence on this principal component. The variable S04 has a loading of -0.056 on PC10, showing a weak negative influence on this principal component. The variable Na has a loading of 0.312 on PC10, showing a moderate positive influence on this principal component. The variable K has a loading of -0.043 on PC10, showing a very weak negative influence on this principal component. The variable Ca has a loading of -0.402 on PC10, showing a moderate negative influence on this principal component. The variable Mg has a loading of 0.110 on PC10, showing a weak positive influence on this principal component. The variable T.H has a loading of -0.177 on PC10, showing a moderate negative influence on this principal component. The variable SAR has a loading of 0.185 on PC10, showing a weak positive influence on this principal component. The variable RSC has a loading of 0.186 on PC10, showing a weak positive influence on this principal component. Based on this we can draw conclusions, this was one way that we analyzed the data.

3. Final PCA Discussion.

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(3.1) would show 2 Components, but it only ended up explain 59% of variance(3.2), 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 to much information. This change was better as I got a better understanding of the data and the new PC variables(3.3). The first PC considered of 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.

Appendix

3.1

A graph of a person with a line graph

Description automatically generated with medium confidence

3.2

A close-up of a computer code

Description automatically generated

3.3

A screenshot of a computer code

Description automatically generated

Code used:

library(dplyr)

library(ggplot2)

df <- read.csv("data.csv")

numeric\_df <- df %>% select(where(is.numeric))

print("Count of missing values in each column:")

print(colSums(is.na(numeric\_df)))

print("Count of infinite values in each column:")

print(colSums(sapply(numeric\_df, is.infinite)))

numeric\_df <- numeric\_df %>%

mutate(across(where(is.numeric), ~ ifelse(is.na(.), mean(., na.rm = TRUE), .)))

numeric\_df[sapply(numeric\_df, is.infinite)] <- NA

numeric\_df <- numeric\_df %>%

mutate(across(where(is.numeric), ~ ifelse(is.na(.), mean(., na.rm = TRUE), .)))

numeric\_df\_scaled <- scale(numeric\_df)

pca\_res <- prcomp(numeric\_df\_scaled, center = TRUE, scale. = TRUE)

summary(pca\_res)

print(pca\_res$rotation)

print(pca\_res$x)

screeplot(pca\_res, type = "lines", main = "Scree Plot")

biplot(pca\_res, main = "PCA Biplot")

plot(pca\_res, type = "l", main = "Explained Variance by Principal Components")

pca\_sd <- pca\_res$sdev

variance\_explained <- (pca\_sd^2) / sum(pca\_sd^2)

scree\_data <- data.frame(PC = 1:length(variance\_explained), VarianceExplained = variance\_explained)

ggplot(scree\_data, aes(x = PC, y = VarianceExplained)) +

geom\_bar(stat = "identity", fill = "steelblue") +

geom\_line(aes(x = PC, y = cumsum(VarianceExplained)), color = "red", size = 1, group = 1) +

geom\_point(aes(x = PC, y = cumsum(VarianceExplained)), color = "red", size = 3) +

labs(title = "Scree Plot", x = "Principal Components", y = "Proportion of Variance Explained") +

theme\_minimal() +

theme(axis.text.x = element\_text(size = 12), axis.text.y = element\_text(size = 12))

pca\_data <- as.data.frame(pca\_res$x[, 1:3])

loadings <- pca\_res$rotation

loadings[, 1:5]

pca\_data <- pca\_res$x[, 1:5]

set.seed(123)

kmeans\_result <- kmeans(pca\_data, centers = 3, nstart = 25)

kmeans\_result$cluster

pca\_data\_with\_clusters <- as.data.frame(pca\_data)

pca\_data\_with\_clusters$Cluster <- as.factor(kmeans\_result$cluster)

ggplot(pca\_data\_with\_clusters, aes(x = PC1, y = PC2, color = Cluster)) +

geom\_point(size = 3) +

theme\_minimal() +

labs(title = "K-means Clustering on PCA Components", x = "PC1", y = "PC2") +

scale\_color\_manual(values = c("red", "green", "blue"))

wcss <- numeric(10)

for (i in 1:10) {

kmeans\_temp <- kmeans(pca\_data, centers = i, nstart = 25)

wcss[i] <- kmeans\_temp$tot.withinss

}

ggplot(data.frame(Clusters = 1:10, WCSS = wcss), aes(x = Clusters, y = WCSS)) +

geom\_line() +

geom\_point() +

labs(title = "Elbow Method for Optimal Clusters", x = "Number of Clusters", y = "WCSS")