

A decorative pattern of hexagons in various colors (light blue, orange, beige, grey, dark blue) arranged in a honeycomb-like structure on the left side of the slide. Some hexagons are solid, while others are outlined.

# Advanced Statistical Computing Final Project

Zeynep Cetin

A single outlined hexagon located at the bottom right of the slide.

# Project 1: Exploratory Data Analysis

## icecream

Only contains two variables: temperature, and ice cream profits and contains 365 observations.

## adultr

Training dataset that includes categories such as age, work class, education level, etc. In total, the dataset contains 15 columns and 36.6 thousand observations.

## abalone

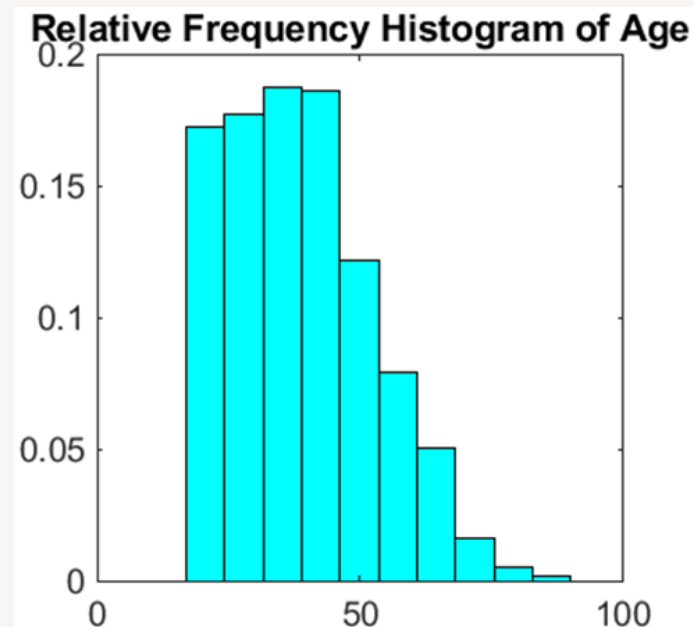
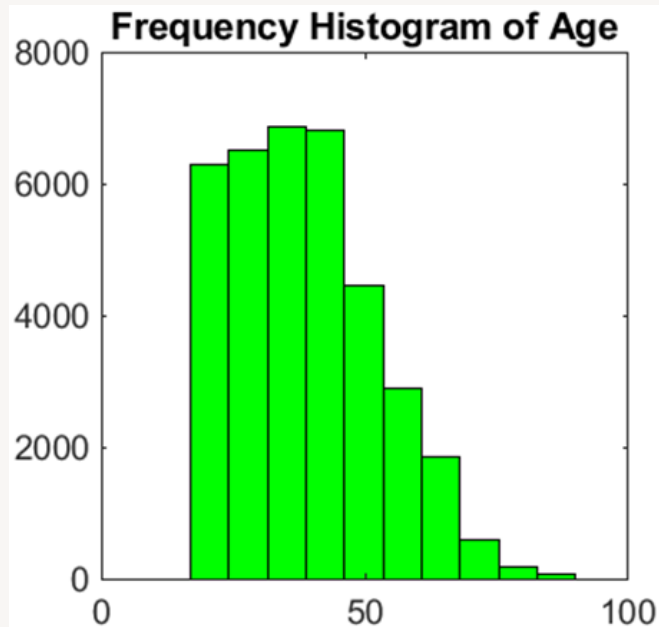
Dataset comes from an original study titled “The Population Biology of Abalone in Tasmania.” Has 9 columns and 4177 observations.



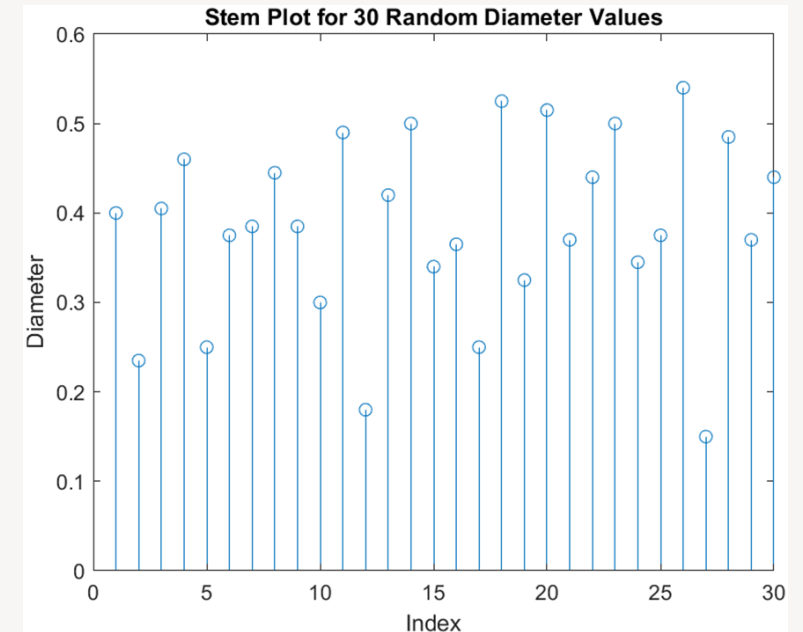
## NYHouse

Dataset contains prices and properties of New York houses; it has 17 variables and 4801 observations.

# Project 1: Exploratory Data Analysis

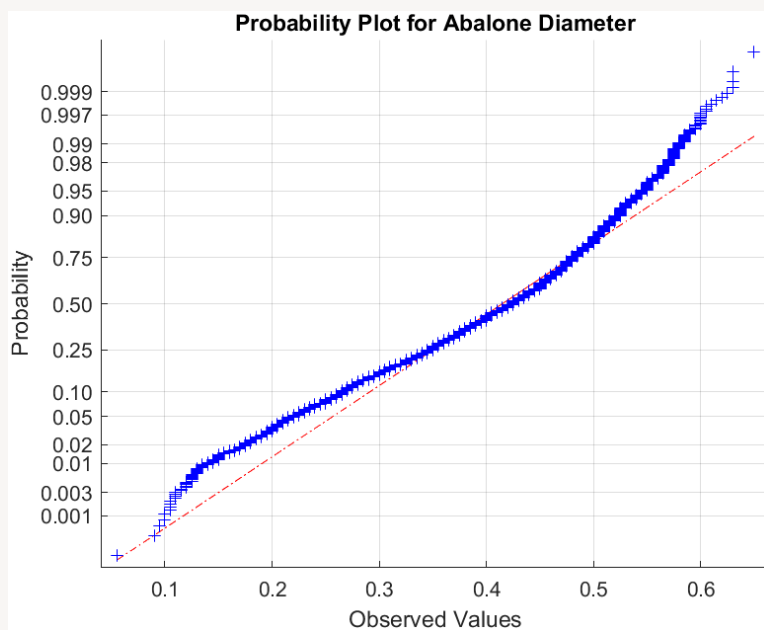


From the histograms, it could be said that the age variable was a bit skewed right.  
The mean age of the dataset was 38.66

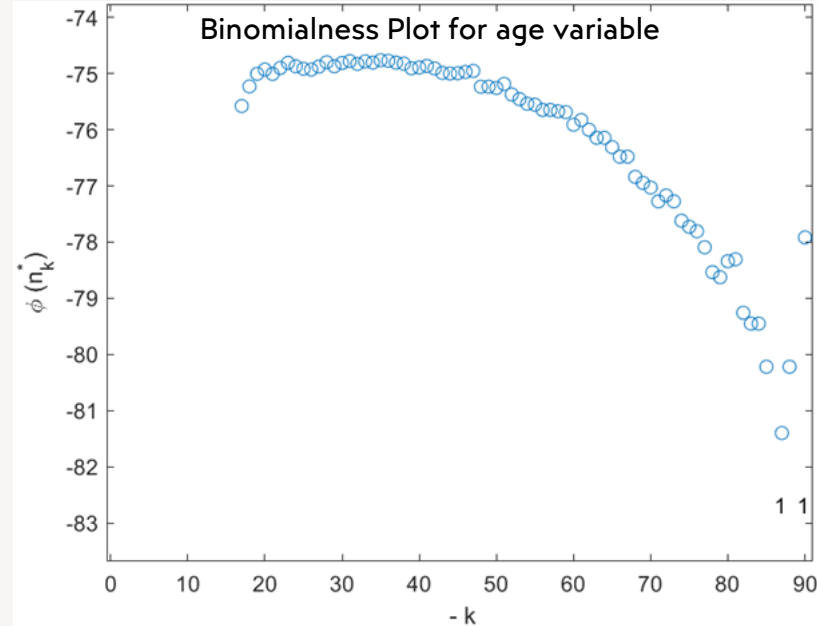


The diameter of the  
abalones varies from 0.15  
to 0.55 for the 30 random  
observations.

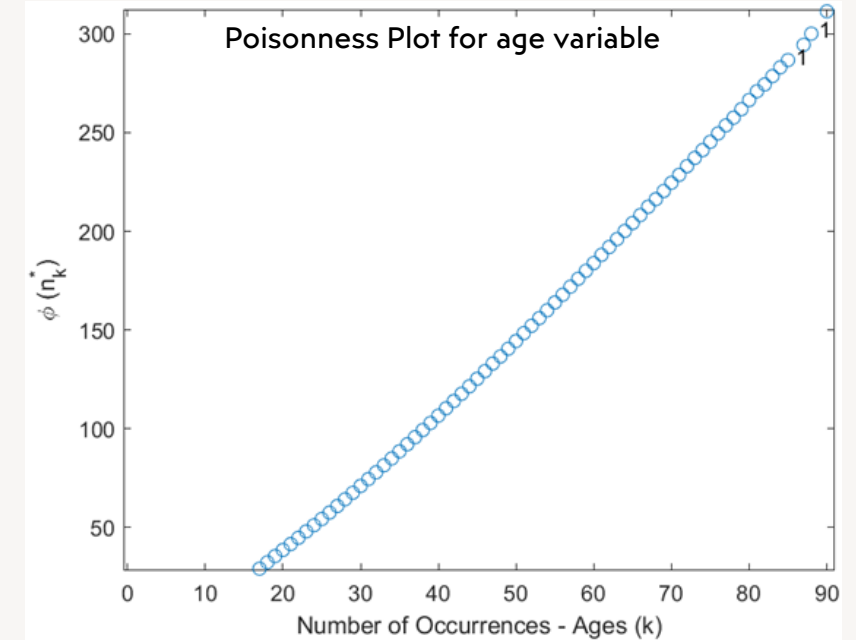
# Project 1: Exploratory Data Analysis



The shape of the line is not very linear, which could mean the data set might not be from a normal distribution.

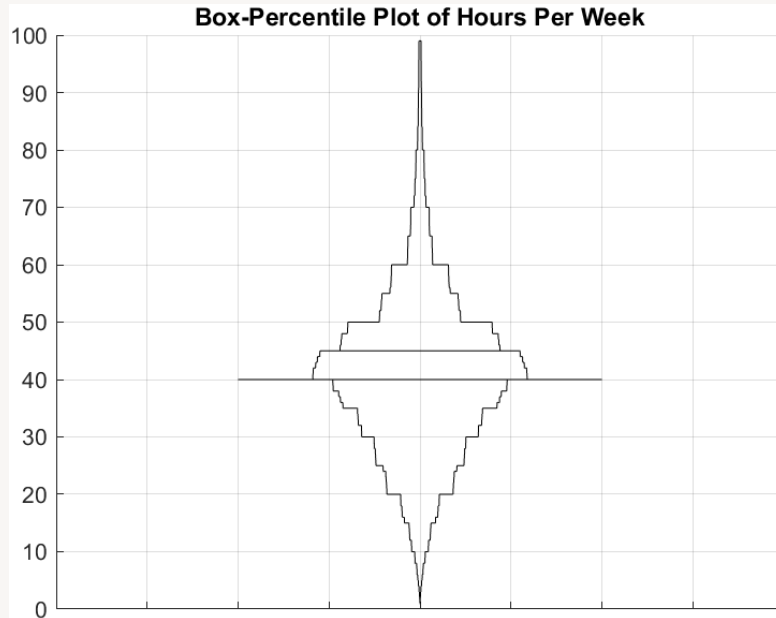


As the relationship is not linear, this indicates that the binomial model is not adequate.

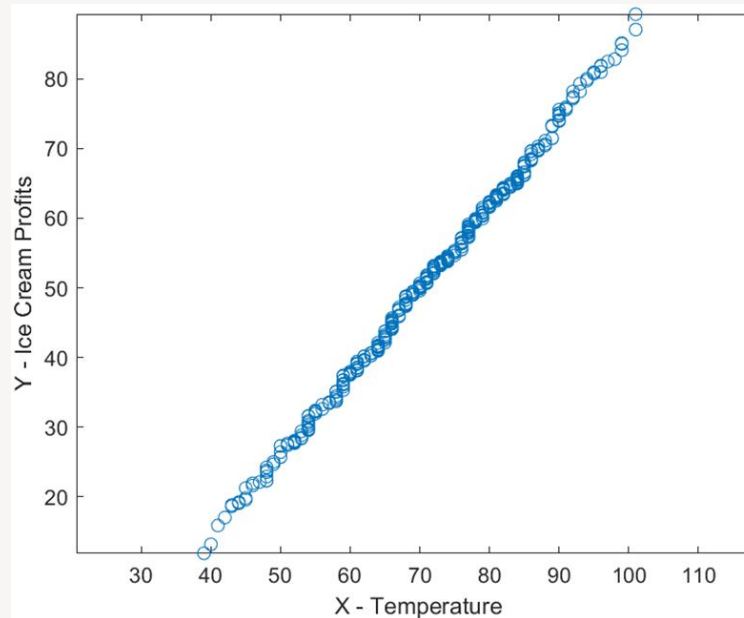


Since the shape follows a straight line, a Poisson distribution is a reasonable model for this data.

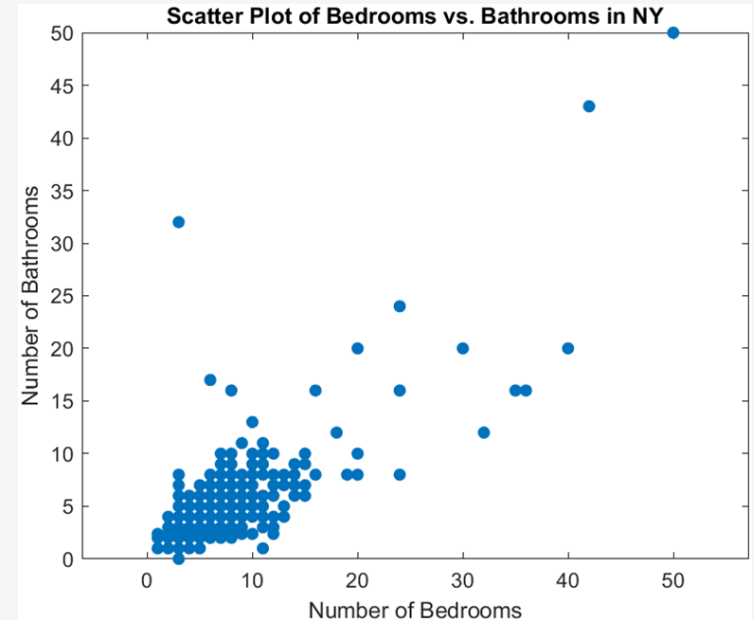
# Project 1: Exploratory Data Analysis



The mean hours per week worked was 40.424, which aligns with the standard 40-hour work week for most Americans

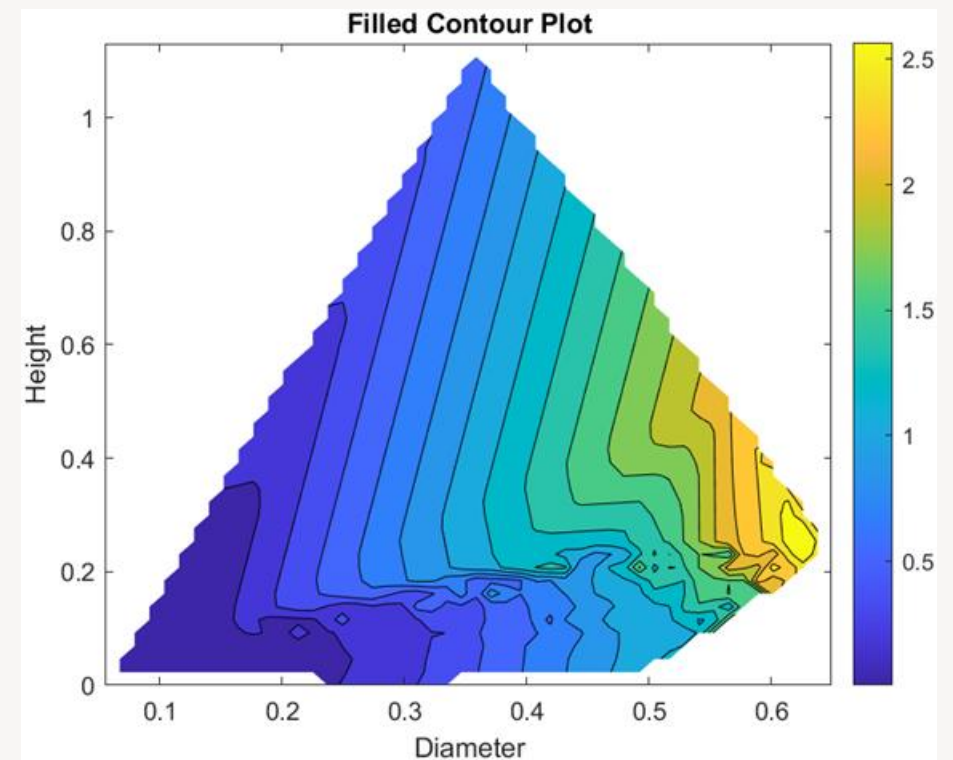
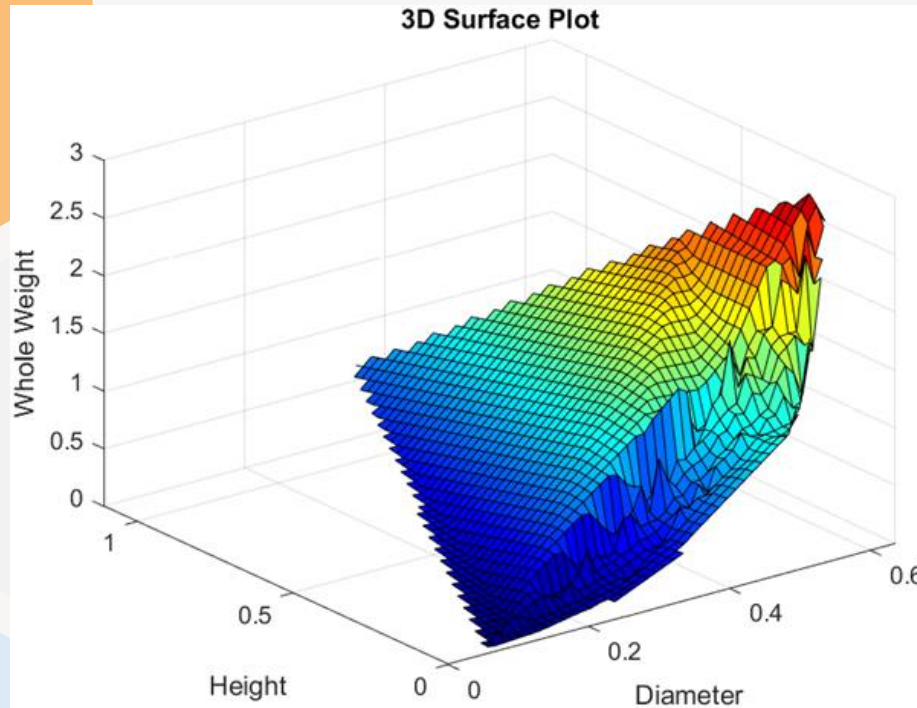


From this Q-Q plot, it can be said that the temperature and the ice cream profits follow the same distribution as it is a linear line.



It can be seen here that most of our data is in between 1 to 10 bathrooms and 1 to 10 bedrooms. There is a linear relationship between the number of bedrooms.

# Project 1: Exploratory Data Analysis



For the abalones, as the diameter and the height of the abalones increase, the weight also seems to increase. The weight is the least when the height and the diameter is the smallest. The diameter contributes more to the weight than the height.

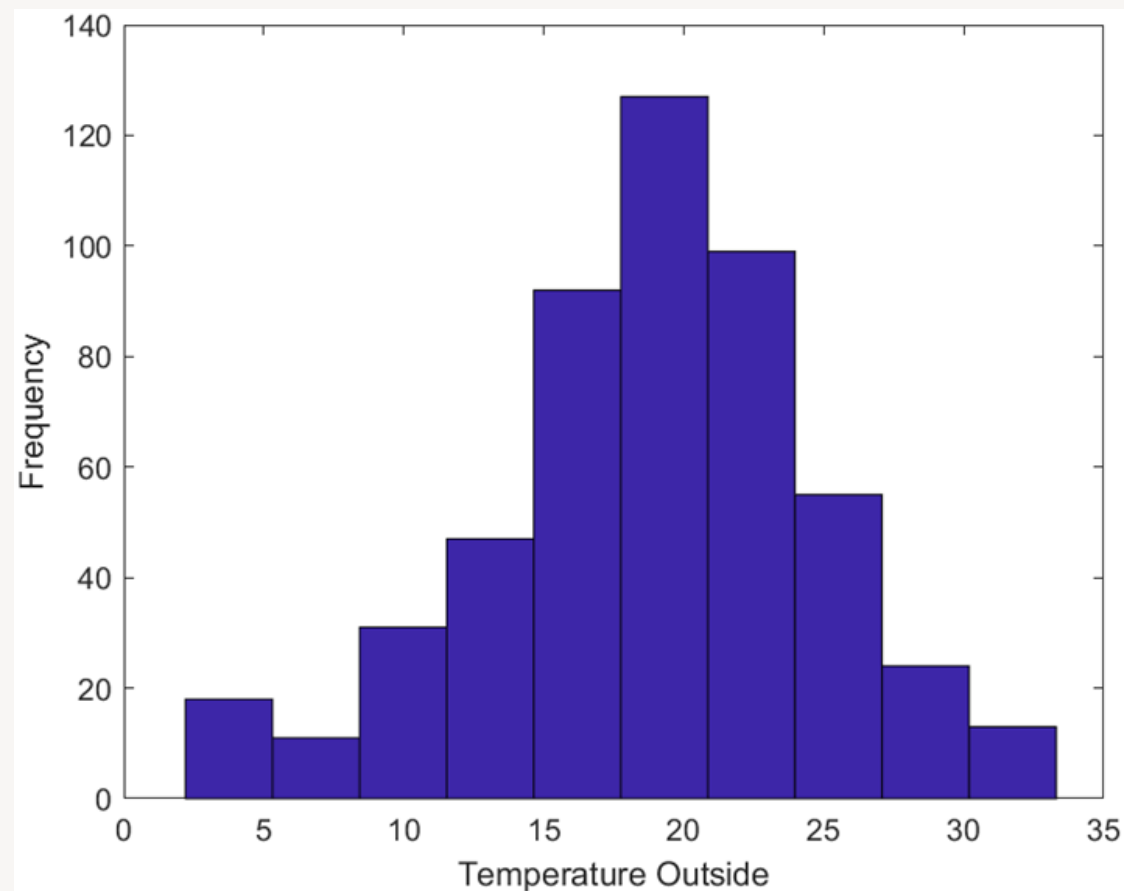
# Project 2: Monte Carlo and Data Partitioning Methods for Inferential Statistics

For this project, the forestfires dataset was obtained from UC Irvine Machine Learning Repository by Cortez and Morais, 2007.

The dataset has 517 observations and 12 variables, but the only used variables were the temperature (temp) and the relative humidity (RH) in this project.

The assumed population mean is 20 and the population standard deviation is 6 for this project.

The data seems to be normally distributed. The mean of the data is 18.9 and a standard deviation of 5.8.



# Project 2: Monte Carlo

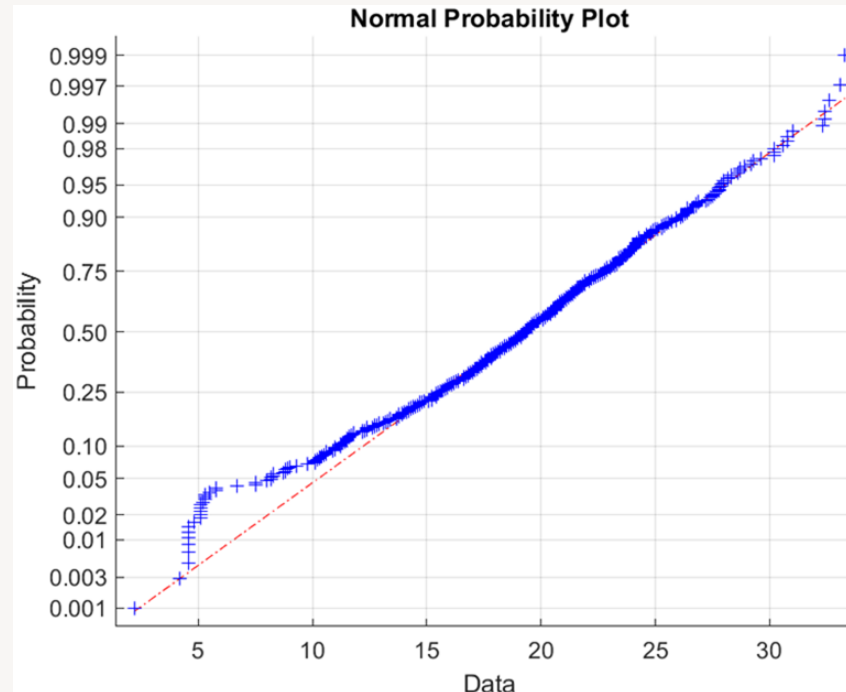
## Monte Carlo Hypothesis Testing by Critical Value

The hypothesis in this project is:

$$H_0: \mu = 20^\circ\text{C}$$


$$H_a: \mu \neq 20^\circ\text{C}$$

The observed value of the test statistic is -4.21. After running 1000 Monte Carlo trials, the estimated critical value is -1.67. Since 4.21 is larger than the critical value, we reject the null hypothesis.



It is assumed that a normal distribution for the data is reasonable from this normal probability plot.

## Monte Carlo Hypothesis Testing (P-value)

 pvalhat 0

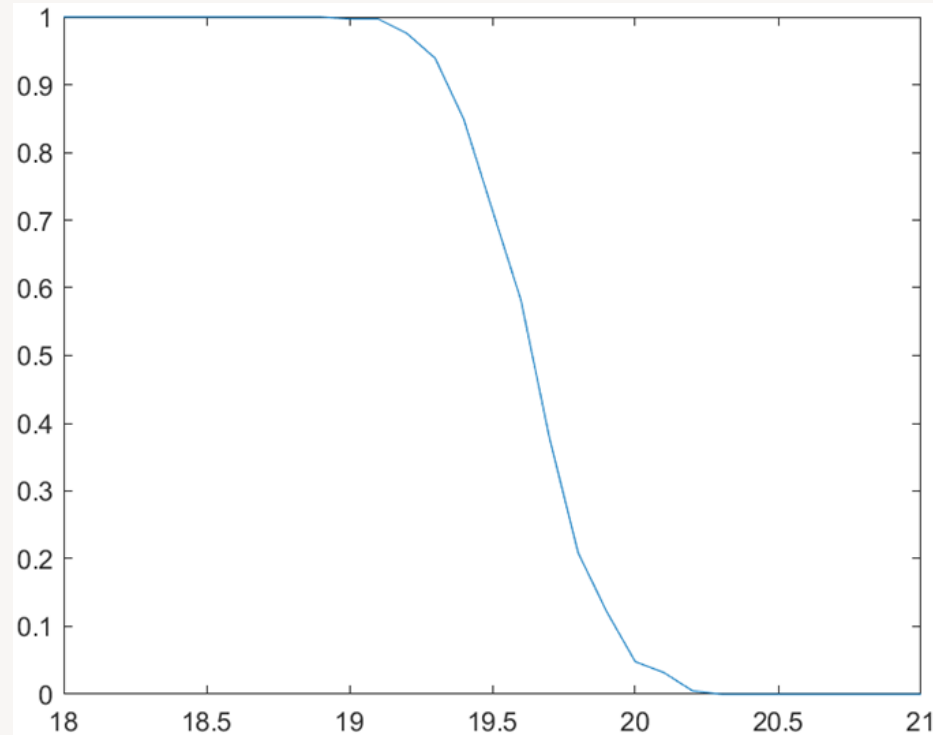
The p-value turned out to be 0, which means that none of the simulated test statistics from the simulation were more extreme than the observed test statistic in the sample.



# Project 2: Monte Carlo

## Monte Carlo Assessment of Hypothesis Testing (Type I & Type II Error)

From the simulation, the estimated value of 0.046 was obtained, which is very close to the desired probability of the Type I error of 0.05 with a critical value of -1.645.



The power (probability of correctly rejecting a false null hypothesis) of the test decreases as the true value for  $\mu$  gets closer to 20.

## Bootstrap Estimate of Bias and Standard Error

sebmat	0.1814
biasb	-0.0013

The standard error and the bias of the standard deviation came out to be 0.1814 and -0.0013, respectively. The bias is smaller than the standard error.

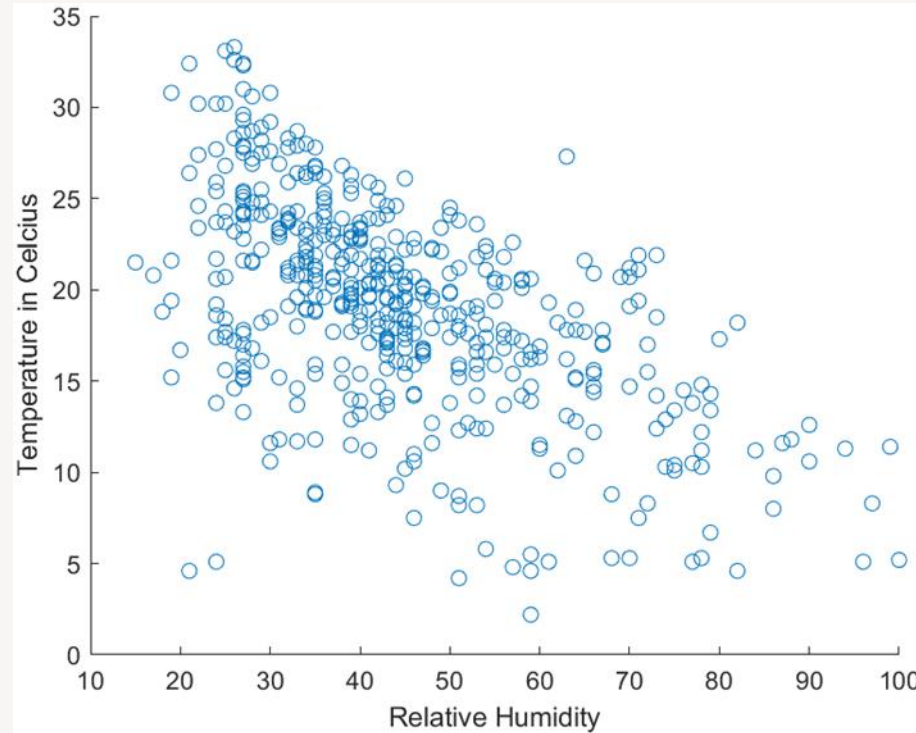
# Project 2: Data Partitioning

## Cross-Validation

In this project, Y denotes the temperature of the day in Celsius, and the X denotes the relative humidity. In this application, linear, quadratic, and cubic models are fit to the data.

pe1	24.4907
pe2	24.5570
pe3	24.5423

The linear fit is the best for this model as it has the lowest prediction error of 24.49



From the scatterplot, it could be said that a linear model is reasonable for the relationship between temperature and humidity.

## Jackknife and Jackknife after Bootstrapping

Jackknife resulted in the standard error of 0.1869 and a bias of -0.003 were observed.

Jackknife after Bootstrapping resulted with the estimate of the standard error of the standard deviation to be 0.1843.

In both cases of Jackknife, the standard errors were higher than the standard error of 0.1814 and a bias of -0.0013 for bootstrapping.

# Project 2: the Confidence Intervals

Bootstrap confidence intervals use resampling to estimate the confidence interval of a statistic, in this case, the standard deviation. The method involves repeatedly resampling the original data with replacement to create bootstrap samples, calculating the statistic of interest for each sample, and then finding the desired percentiles of the bootstrap distribution to establish the confidence interval.

	Lower Value	Higher Value
Bootstrap Standard CI	5.513	6.100
Bootstrap-t CI	5.508	6.172
Bootstrap Percentile CI	5.48	6.104
Better Bootstrap CI	5.516	6.110

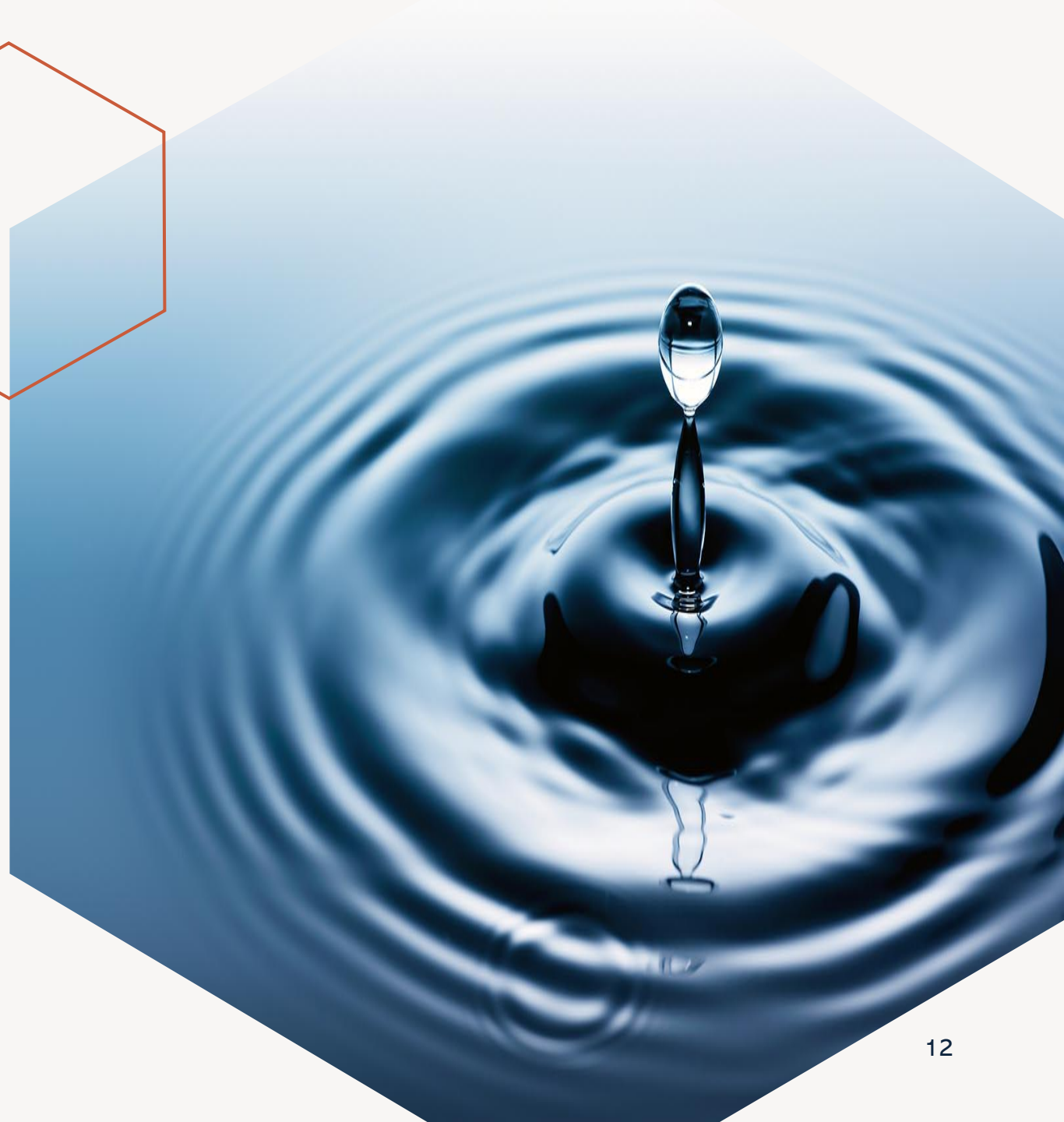
The Bootstrap Standard CI had the lowest range, but the ranges were very close to each other.  
Bootstrap-t Confidence Interval has the largest range.

# Project 3: Data Analysis Using Supervised Learning and Unsupervised Learning

The waterpotability dataset was used, which contains 10 variables: pH, Hardness, Solids, Chloramines, Sulfate, Conductivity, Organic Carbon, Trihalomethanes, Turbidity, and Potability.

Potability variable is the target variable where 1 represents potable water and 0 represents not potable water.

Observations with missing values were removed to simplify data preprocessing. While this approach may lead to a loss of data, it was chosen for the sake of focusing on other aspects of the project. Alternative methods, such as imputation, could also have been considered for handling missing values.



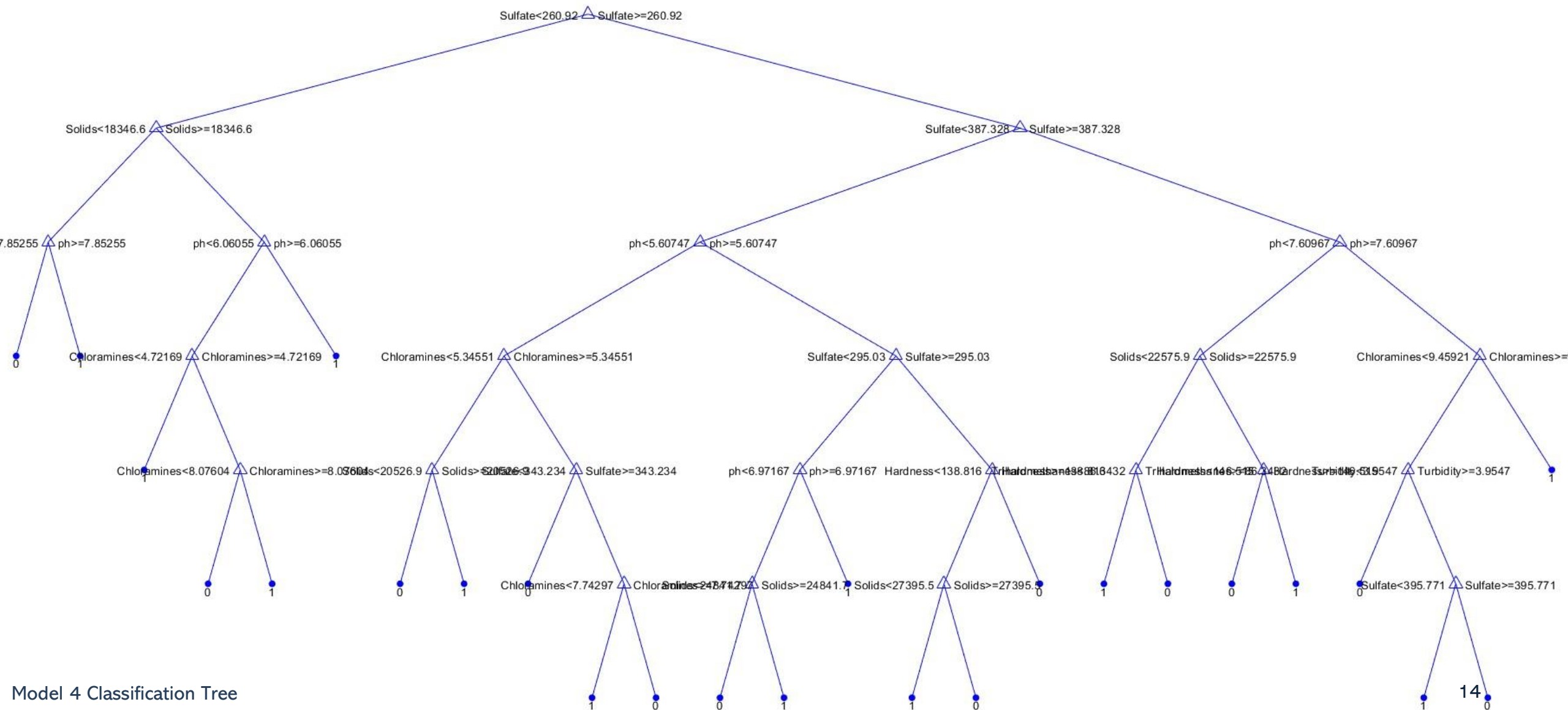
# Project 3: Supervised Learning – Classification Trees

Classification trees are a type of machine learning model which is used to predict categorical labels based on input features. They work by splitting the data into subsets based on feature values.

Model Name	Specifications	Resubstitution Loss	k-fold Loss
Model 1	none	0.06713	0.41323
Model 2	Max number of splits 20	0.31079	0.36698
Model 3	Max number of splits 30	0.30134	0.363
Model 4	Max number of splits 30, Min leaf size 4, GDI split criterion	0.30433	0.35554
Model 5	Max number of splits 30, Min leaf size 4, deviance split criterion	0.31576	0.36101
Model 6	Optimize Hyperparameters – Max number of splits 773, Min leaf size 48, split criterion GDI	0.29687	0.38886

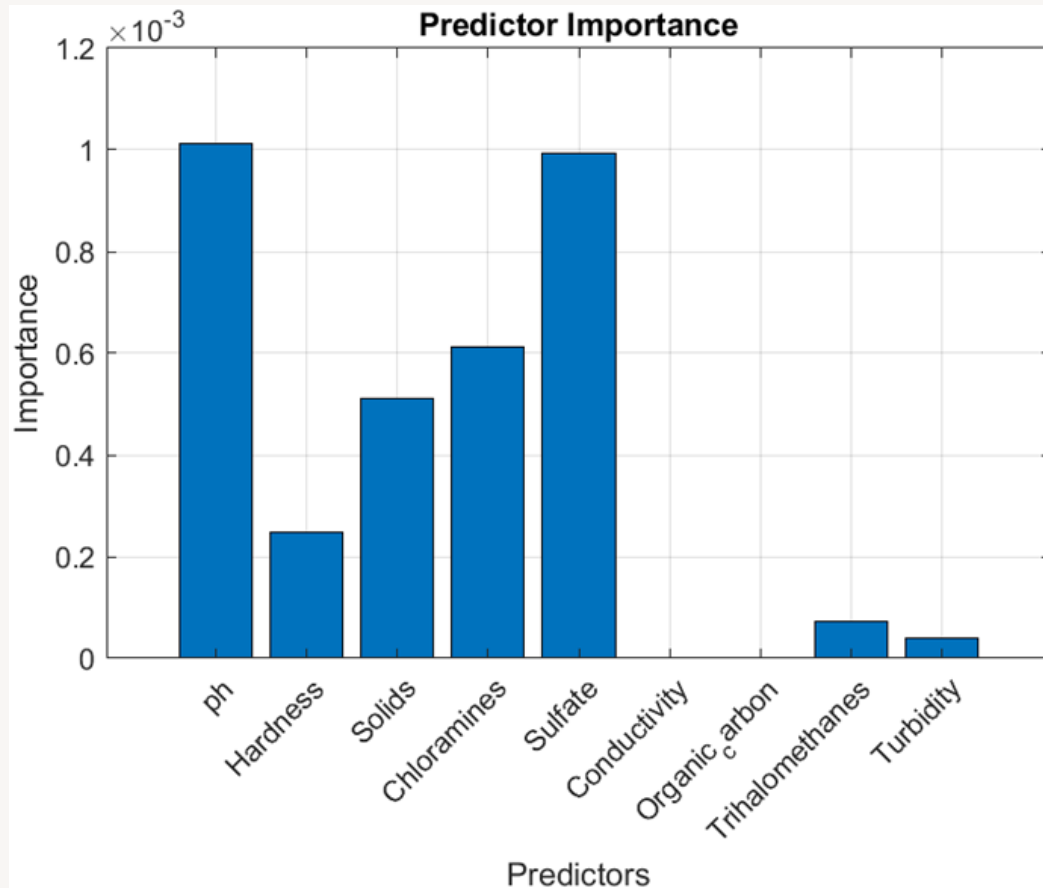
From the table, model 4 seems like the best choice among the models as it provides the best balance of performance on unseen data (lowest k-fold loss), while avoiding overfitting on the training data. Model 1 seems like the poorest choice, as the resubstitution loss is very low, indicating overfitting.

# Project 3: Supervised Learning – Classification Trees



Model 4 Classification Tree

# Project 3: Supervised Learning – Classification Trees



Predictor Importance was run using model 4. The top 3 most important predictors were pH, sulfate, and chloramines.

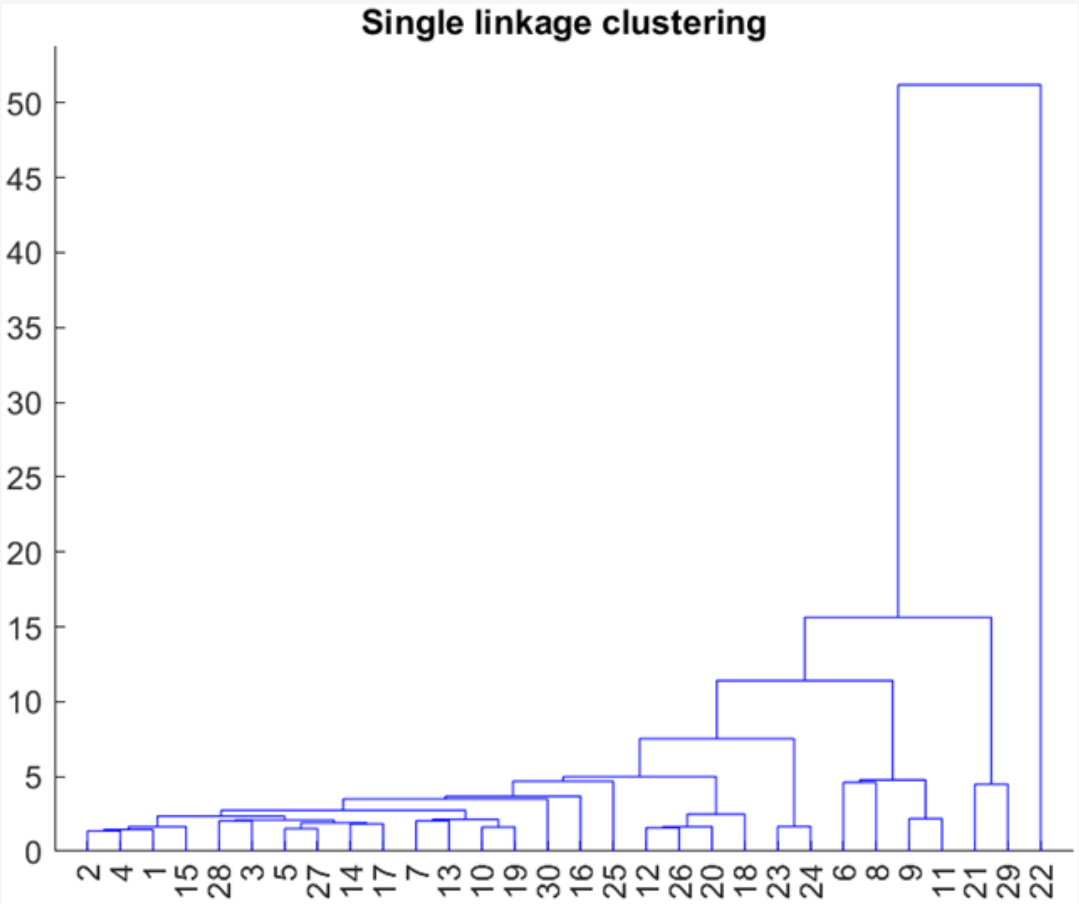
Conductivity and organic carbon were not important predictors to the model.

Additionally, Trihalomethanes and Turbidity were very low in importance.

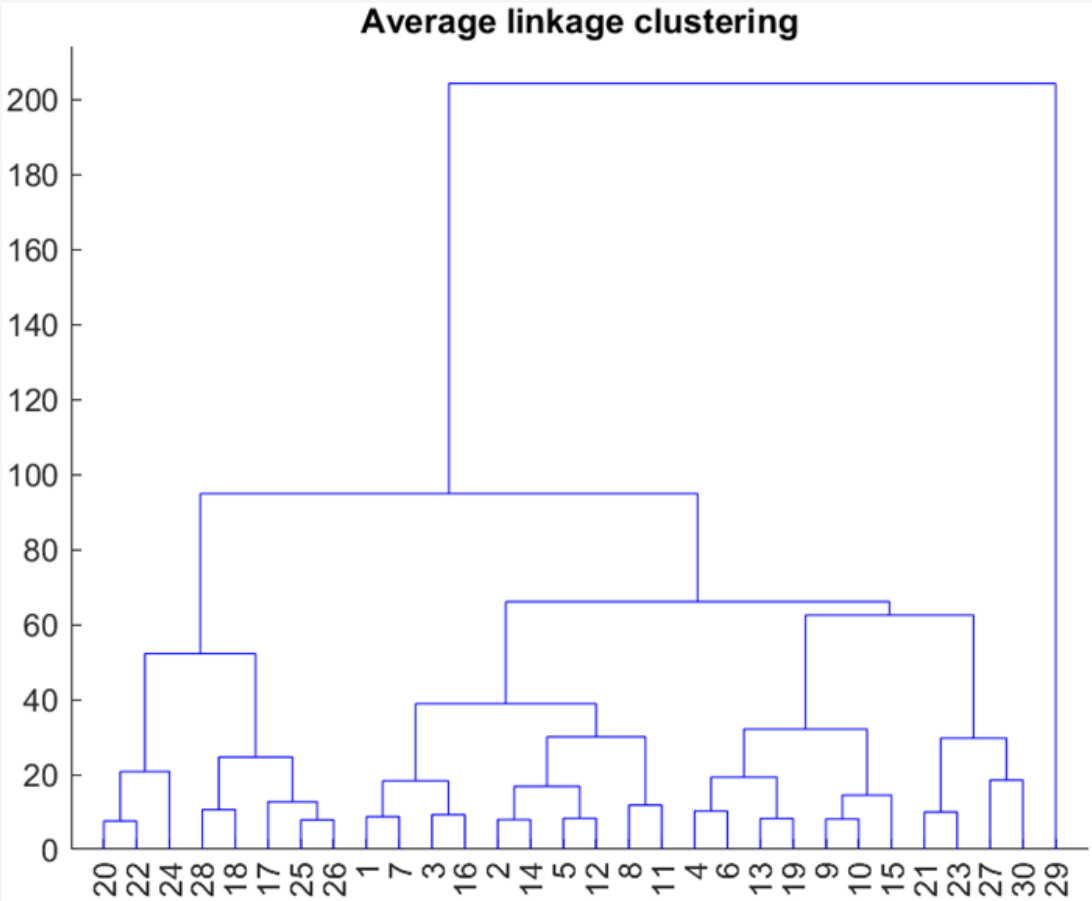


# Project 3: Unsupervised Learning – Agglomerative Hierarchical Clustering

Agglomerative hierarchical clustering was demonstrated using the “Sulphate” variable from the same dataset used as the supervised learning dataset.



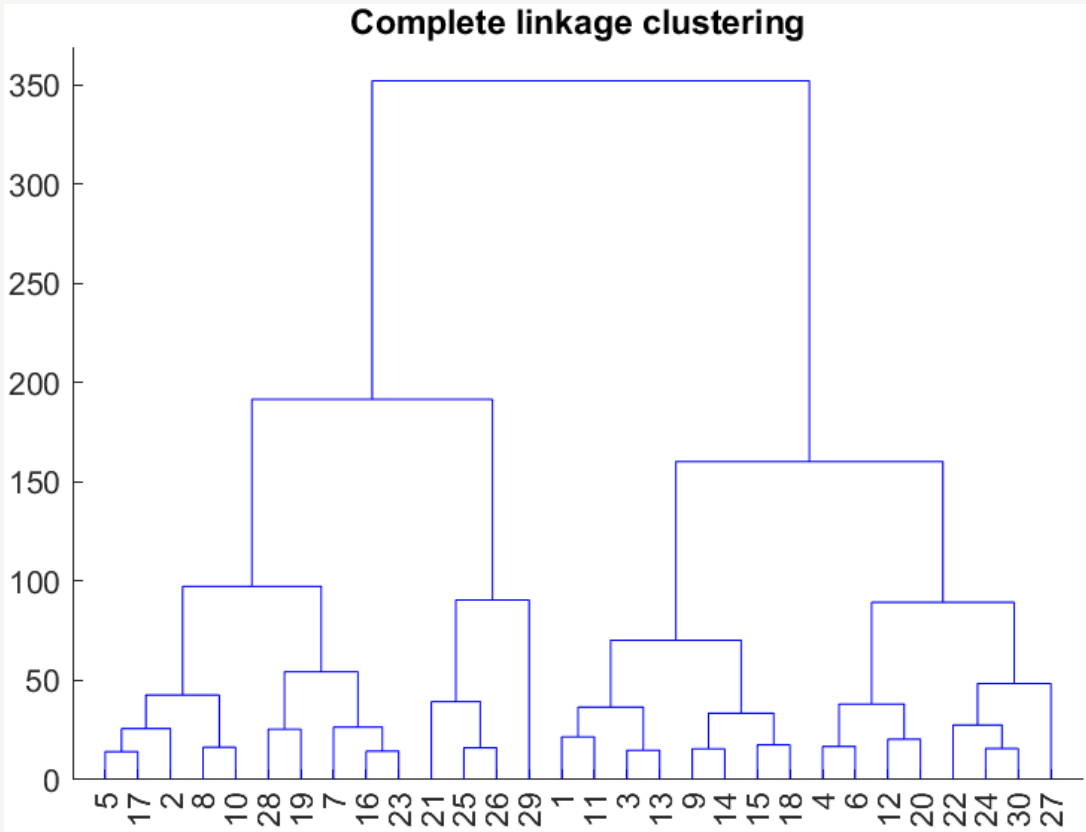
Single Linkage: the distance between clusters is defined as the shortest distance between any two data points in different clusters.



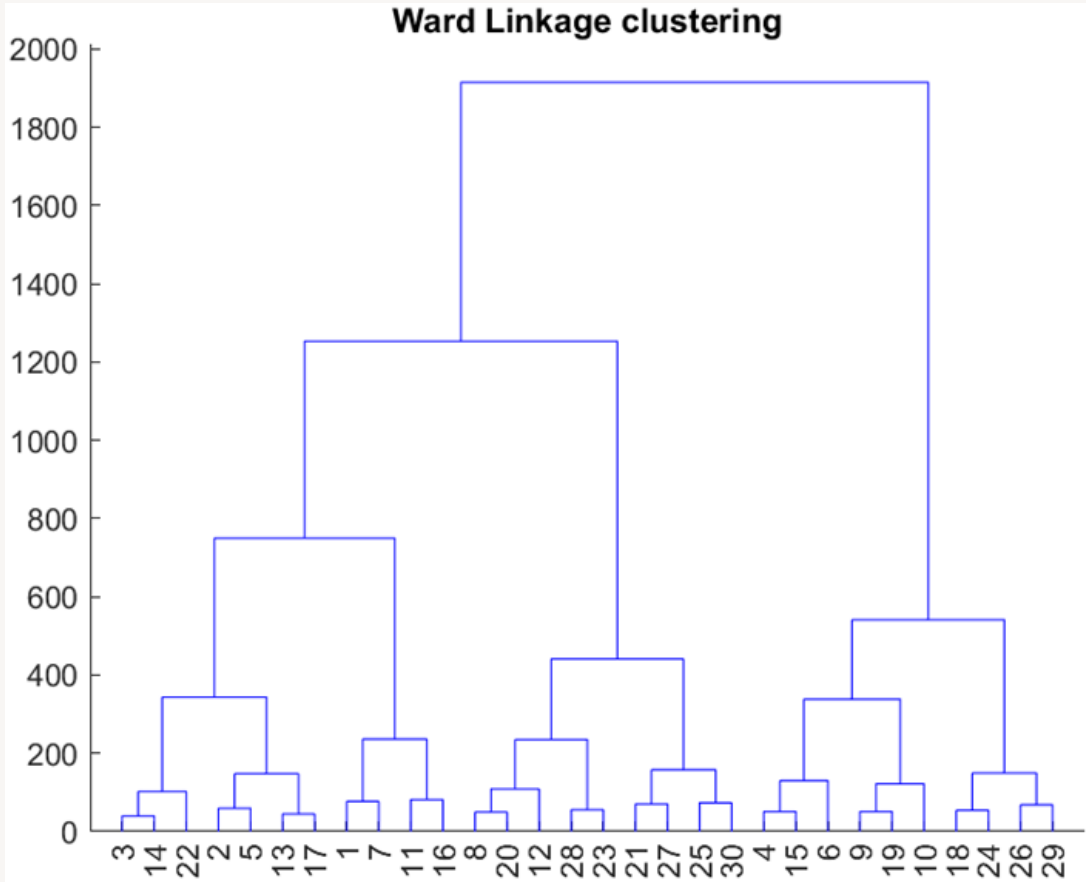
Average Linkage: the distance between clusters is defined as the average distance between all pairs of points.



# Project 3: Unsupervised Learning – Agglomerative Hierarchical Clustering



Complete Linkage: the distance between clusters is defined as the maximum distance between any two data points in different clusters.



Ward's Linkage: the distance between clusters is determined by the increase in the sum of squared distances when merging two clusters.

# Project 3: Unsupervised Learning – Agglomerative Hierarchical Clustering

## Cophenetic Correlation Coefficients

Single Linkage	0.43125
Complete Linkage	0.5659
Average Linkage	0.73109
Ward's Linkage	0.65299

A high coefficient indicates that the linkage accurately represents the data. The highest cophenetic coefficient is from the average linkage, making it a better choice than the others.



**Thank  
You!**

# References:

## Datasets:

- <https://www.kaggle.com/datasets/uom190346a/water-quality-and-potability>
- <https://www.kaggle.com/datasets/hemanthpingali/adult-census-dataset>
- <https://www.kaggle.com/datasets/ryanholbrook/dl-course-data>
- <https://www.kaggle.com/datasets/raphaelmanayon/temperature-and-ice-cream-sales>
- <https://www.kaggle.com/datasets/nelgiriyeewithana/new-york-housing-market>
- <https://archive.ics.uci.edu/dataset/162/forest+fires>

## Images:

- TROUT55/GETTY IMAGES
- <https://www.theguardian.com/environment/2018/aug/19/poachers-abalone-south-africa-seafood-divers>

