Part 1A:

**Measures of Central Tendency**

Mode function:

Text

Description automatically generated with medium confidence

Data Preprocessing and stats: Applied to each column individually ie. EPI, DALY, AIR\_E, CLIMATE

Ex:

Text

Description automatically generated

Mean:

EPI = 58.37379

DALY = 53.94347

AIR\_E = 51.2915

CLIMATE = 56.50382

Median:

EPI = 59.24609

DALY = 60.34839

AIR\_E = 50.45814

CLIMATE = 56.32458

Mode:

EPI = 36.34838

DALY = 86.8625

AIR\_E = 33.12662

CLIMATE = 53.84652

Part 1B:

**Regression**

Created a separate dataframe, EPI\_DATA\_subset, containing only "EPI\_regions", "EPI", "DALY", "AIR\_E", "CLIMATE" as the columns.

Standardize the EPI, DALY, AIR\_E, CLIMATE columns to have mean 0 and variance 1. This is to obtain meaningful regression coefficients

Snippet:

EPI\_subset\_standardize = EPI\_DATA\_subset

EPI\_subset\_standardize[c("EPI", "DALY", "AIR\_E", "CLIMATE")] = scale(EPI\_subset\_standardize[c("EPI", "DALY", "AIR\_E", "CLIMATE")])

Generated linear regression models for sub-datasets based on 4 EPI\_regions, looked at largest coefficient:

For example:

EPI\_region == Sub-Saharan Africa

Snippet:

mm = lm(EPI~DALY+AIR\_E+CLIMATE

, EPI\_subset\_standardize[EPI\_subset\_standardize$EPI\_regions == "Sub-Saharan Africa",])

coef(mm)

nrow(EPI\_subset\_standardize[EPI\_subset\_standardize$EPI\_regions == "Sub-Saharan Africa",])

#DALY has highest influence

Results:

Sub-Saharan Africa:

(Intercept) DALY AIR\_E CLIMATE

0.13783438 0.99063359 -0.03993663 0.40830849

Eastern Europe and Central Asia:

(Intercept) DALY AIR\_E CLIMATE

0.10572167 0.95986111 0.04966672 0.58295451

Middle East and North Africa:

(Intercept) DALY AIR\_E CLIMATE

-0.3840508 1.0460516 0.2875852 0.7221674

North America: (Not the most reliable result since the are only 2 data rows for North America)

(Intercept) DALY AIR\_E CLIMATE

-0.3858950 0.8057633 NA NA

Appears as though DALY has the most influence on EPI since it has the largest coefficient (largest weight)

Part 2:

Exercise 1:

A picture containing text

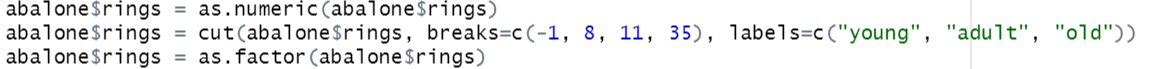
Description automatically generated

ROLL\_predicted = 81437.04

ROLL\_predicted = 137452.6

Exercise 2:

Encode rings column as (young, adult, old)

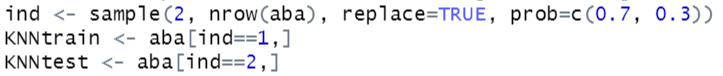


Standardize data for columns 1:7

Text

Description automatically generated with medium confidence

Obtain training and testing datasets from dataframe



For multiple nearest neighbor values compute accuracy metric and cross table

Ex: k = 30

Text

Description automatically generated

Table

Description automatically generated

Results:

K = 30

Table

Description automatically generated

Accuracy = 0.677116

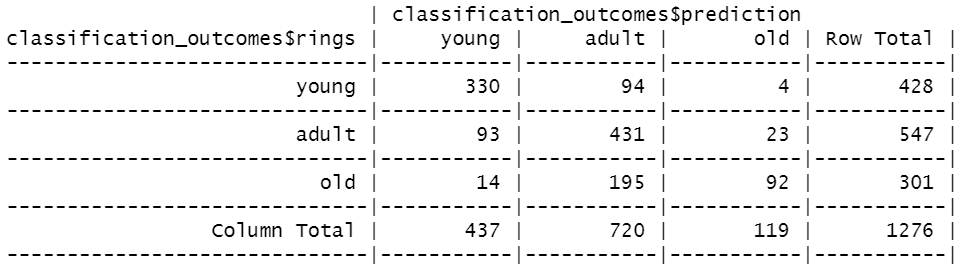
K = 55

Table

Description automatically generated

Accuracy = 0.6677116

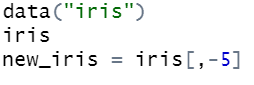
K = 70



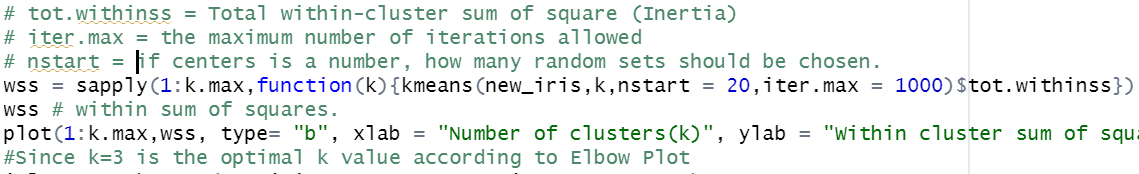
Accuracy = 0.6684953

Exercise 3:

Create new dataframe without 5th column:



Generating Elbow plot to find optimal clusters (k):



Based off Elbow plot, k=3 is the best value. Results:

