Imputation and EDA

Ifechi Umeh

4/11/2022

data <- read\_csv("Data Clean/plastic\_1.csv")  
  
data <- data %>% replace\_with\_na(replace = list(gdp\_per\_capita = "no data")) # some values labelled "no data"  
sum(is.na(data$gdp\_per\_capita)) #Number of NAs: 50; proportion missing: 0.204918  
  
data$gdp\_per\_capita <- as.numeric(data$gdp\_per\_capita)  
  
data <- select(data, -c("Entity"))  
  
names(data) <- c("waste\_gen", "population", "growth\_rate", "area", "density", "fert\_rate", "life\_exp", "mort\_rate", "per\_cap\_waste", "misman\_waste", "share\_misman", "per\_cap\_gdp")

summary(data)  
  
md.pattern(data) #visualizing missing values  
  
vis\_miss(data, cluster = F, sort\_miss = T, show\_perc = T,  
 show\_perc\_col = F, large\_data\_size = 9e+05)

imputed\_data <- mice(data,   
 m = 31, # m selected using max(5, 100p) where p=0.3114754 (highest prop. of NAs)   
 method = 'rf', # uses ranger package  
 seed = 79069084) # seed selected using sample(1:999999999, size = 1)

Error from using “pmm” method for imputation: “Error in solve.default(xtx + diag(pen)) : system is computationally singular: reciprocal condition number = 6.10832e-20”

table\_10 <- complete(imputed\_data, 10) # sample(1:31, size=1)  
summary(table\_10)

# correlation matrix  
cor(data[,unlist(lapply(data, is.numeric))])