plastic land to sea

go

25/2/2020

## R Markdown

### There is a sum up at the end of the document ###

library("caret")

library('dplyr')

#install.packages('tidyverse')  
library(tidyverse)

#install.packages('SnowballC')  
#library(SnowballC)  
#install.packages("SentimentAnalysis")  
#library(SentimentAnalysis)  
library(dplyr)  
library("readxl")  
library(ggplot2)  
library(corrplot)

## corrplot 0.84 loaded

setwd('C:\\Users\\User\\Documents\\ds masters\\sem2\\rnd\\dsets\\plast\_waste\_land')  
plast\_wl <- read\_excel("Plastic\_land\_ocean.xlsx")  
ncol(plast\_wl)

## [1] 14

names(plast\_wl)

## [1] "Country"   
## [2] "Economic status1"   
## [3] "Coastal population2"   
## [4] "Waste generation rate [kg/person/day]3"   
## [5] "% Plastic in waste stream4"   
## [6] "% Inadequately managed waste5"   
## [7] "% Littered waste6"   
## [8] "Waste generation [kg/day]7"   
## [9] "Plastic waste generation [kg/day]7"   
## [10] "Inadequately managed plastic waste [kg/day]7"   
## [11] "Plastic waste littered\r\n [kg/day]7"   
## [12] "Mismanaged plastic waste [kg/person/day]7"   
## [13] "Mismanaged plastic waste in 2010\r\n [tonnes]7"  
## [14] "Mismanaged plastic waste in 2025\r\n [tonnes]7"

# str(plast\_wl)

plasticW <- plast\_wl  
  
names(plasticW)[names(plasticW) == "Mismanaged plastic waste in 2010\r\n [tonnes]7"] <- "Mismanaged\_PL\_2010\_ton"  
names(plasticW)[names(plasticW) == "Plastic waste generation [kg/day]7"] <- "Pl\_waste\_gen\_kgPerday"  
names(plasticW)[names(plasticW) == "Inadequately managed plastic waste [kg/day]7"] <- "Inad\_man\_pl\_waste\_kgPerday"  
names(plasticW)[names(plasticW) == "Waste generation [kg/day]7"] <- "Waste\_gen\_kgPerday"  
names(plasticW)[names(plasticW) == "Economic status1"] <- "Ec\_status"  
names(plasticW)[names(plasticW) == "Coastal population2"] <- "Coastal\_pop"  
names(plasticW)[names(plasticW) == "% Inadequately managed waste5"] <- "InadeqManWaste\_Percent"  
names(plasticW)[names(plasticW) == "% Littered waste6"] <- "LitteredWaste\_Percent"  
names(plasticW)[names(plasticW) == "% Plastic in waste stream4"] <- "Pl\_in\_waste\_stream\_Percent"  
names(plasticW)[names(plasticW) == "Mismanaged plastic waste in 2025\r\n [tonnes]7"] <- "Mismanaged\_PL\_2025\_ton"  
names(plasticW)[names(plasticW) == "Waste generation rate [kg/person/day]3"] <- "Waste\_genRate\_kgPerpersonPerday"  
names(plasticW)[names(plasticW) == "Plastic waste littered\r\n [kg/day]7"] <- "Pl\_wast\_littered\_kgPerday"  
names(plasticW)[names(plasticW) == "Mismanaged plastic waste [kg/person/day]7"] <-"Mismanaged\_PL\_kgPerpersonPerday"  
  
# str(plasticW)

# head(plasticW)

plasticW <- plasticW[complete.cases(plasticW), ]  
#plasticW$Ec\_status <- as.factor(plasticW$Ec\_status)  
plasticW <- plasticW[order(-plasticW$Mismanaged\_PL\_2010\_ton),]  
# head(plasticW)

mydata <- c("Country",  
 "Ec\_status",  
 "Coastal\_pop",  
 "Waste\_gen\_kgPerday",  
 "Pl\_waste\_gen\_kgPerday",  
 "Inad\_man\_pl\_waste\_kgPerday",  
 "Mismanaged\_PL\_2010\_ton",  
 "Mismanaged\_PL\_2025\_ton")  
  
df <- plasticW[mydata]   
# str(df)

...

# head(df)

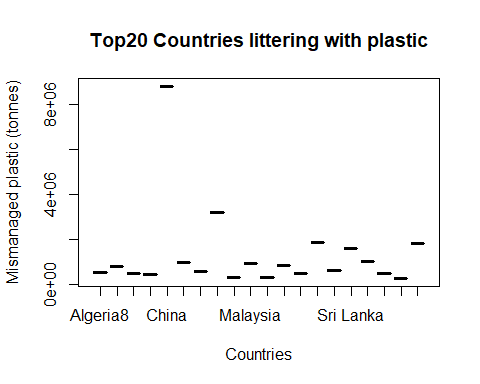
nrow(df)

## [1] 192

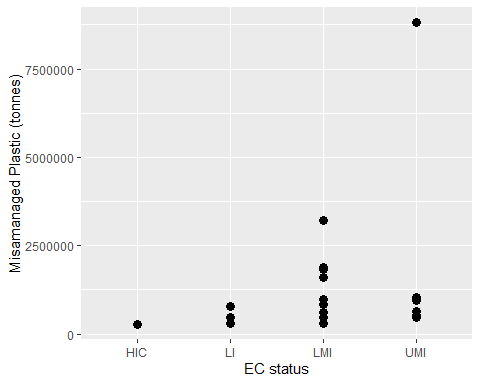
colnames(df)

## [1] "Country" "Ec\_status"   
## [3] "Coastal\_pop" "Waste\_gen\_kgPerday"   
## [5] "Pl\_waste\_gen\_kgPerday" "Inad\_man\_pl\_waste\_kgPerday"  
## [7] "Mismanaged\_PL\_2010\_ton" "Mismanaged\_PL\_2025\_ton"

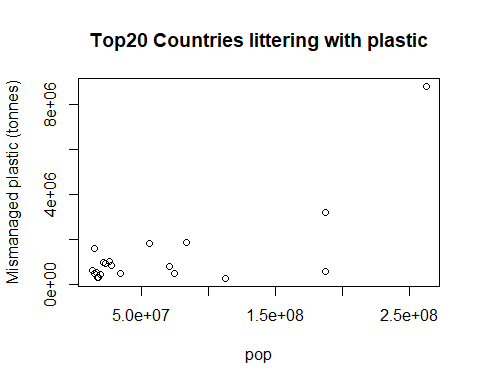
# EDA  
# top 20 countries in mismanaged plastic for 2010 (tonnes)  
dfTop <- df[1:20,]  
  
# plotting top 20 countries and mismanaged plastic  
dfTop$Country2 <- as.factor(dfTop$Country)  
plot(dfTop$Country2, dfTop$Mismanaged\_PL\_2010\_ton, main= 'Top20 Countries littering with plastic',  
 xlab='Countries', ylab='Mismanaged plastic (tonnes)')



# plotting mismanaged plastic and economic status of countries  
plasticW$Ec\_status <- as.factor(plasticW$Ec\_status)  
# low to medium income countries pollute the most.  
qplot(dfTop$Ec\_status, dfTop$Mismanaged\_PL\_2010\_ton, data=dfTop,  
 size=I(3), xlab="EC status", ylab="Misamanaged Plastic (tonnes)")



# population seems to play a role, but no clear indication  
plot(dfTop$Coastal\_pop, dfTop$Mismanaged\_PL\_2010\_ton, main= 'Top20 Countries littering with plastic',  
 xlab='pop', ylab='Mismanaged plastic (tonnes)')



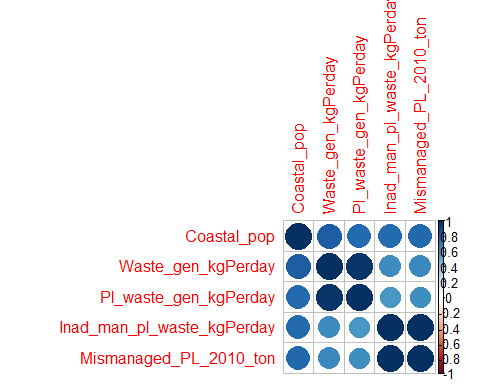
### Creating linear regression models to fit data.  
## lm-1   
  
dfN <- df[,-c(1,2,8)]  
str(dfN)

## Classes 'tbl\_df', 'tbl' and 'data.frame': 192 obs. of 5 variables:  
## $ Coastal\_pop : num 2.63e+08 1.87e+08 8.34e+07 5.59e+07 1.46e+07 ...  
## $ Waste\_gen\_kgPerday : num 2.89e+08 9.74e+07 4.17e+07 4.41e+07 7.43e+07 ...  
## $ Pl\_waste\_gen\_kgPerday : num 31665388 10660505 6237653 5714578 5163689 ...  
## $ Inad\_man\_pl\_waste\_kgPerday: num 23530300 8600093 5035956 4909870 4256120 ...  
## $ Mismanaged\_PL\_2010\_ton : num 8819717 3216856 1883659 1833819 1591179 ...

cor(dfN)

## Coastal\_pop Waste\_gen\_kgPerday Pl\_waste\_gen\_kgPerday  
## Coastal\_pop 1.0000000 0.8276240 0.7709213  
## Waste\_gen\_kgPerday 0.8276240 1.0000000 0.9782371  
## Pl\_waste\_gen\_kgPerday 0.7709213 0.9782371 1.0000000  
## Inad\_man\_pl\_waste\_kgPerday 0.7707409 0.6210473 0.5865949  
## Mismanaged\_PL\_2010\_ton 0.7836694 0.6464420 0.6137701  
## Inad\_man\_pl\_waste\_kgPerday Mismanaged\_PL\_2010\_ton  
## Coastal\_pop 0.7707409 0.7836694  
## Waste\_gen\_kgPerday 0.6210473 0.6464420  
## Pl\_waste\_gen\_kgPerday 0.5865949 0.6137701  
## Inad\_man\_pl\_waste\_kgPerday 1.0000000 0.9994228  
## Mismanaged\_PL\_2010\_ton 0.9994228 1.0000000

corrplot(cor(dfN))

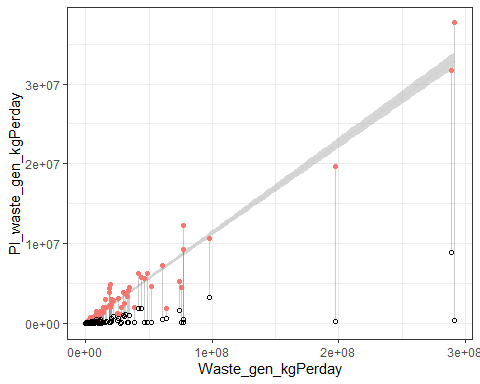


# High correlation between "Waste\_gen\_kgPerday" and "Pl\_waste\_gen\_kgPerday" -> Plastic is main waste  
# High correlation between "Inad\_man\_pl\_waste\_kgPerday" and "Mismanaged\_PL\_2010\_ton " & "Mismanaged\_PL\_2025\_ton"  
myModel=lm(dfN$Mismanaged\_PL\_2010\_ton ~., data=dfN)  
dfN$predicted <- predict(myModel)  
dfN$residuals <- myModel$residuals  
summary(myModel)

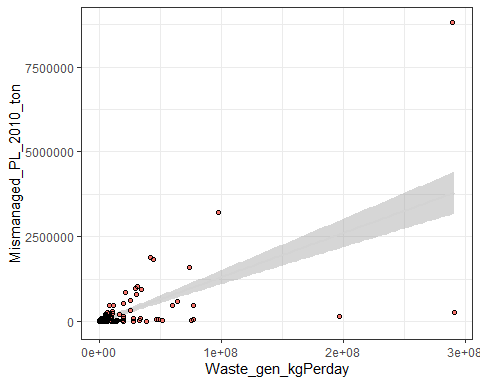
## Warning in summary.lm(myModel): essentially perfect fit: summary may be  
## unreliable

##   
## Call:  
## lm(formula = dfN$Mismanaged\_PL\_2010\_ton ~ ., data = dfN)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.694e-09 -3.100e-11 2.510e-11 8.310e-11 3.242e-09   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -3.698e-10 3.127e-11 -1.183e+01 < 2e-16 \*\*\*  
## Coastal\_pop 4.133e-17 2.215e-18 1.866e+01 < 2e-16 \*\*\*  
## Waste\_gen\_kgPerday -2.151e-17 4.784e-18 -4.497e+00 1.21e-05 \*\*\*  
## Pl\_waste\_gen\_kgPerday 7.300e-03 3.634e-17 2.009e+14 < 2e-16 \*\*\*  
## Inad\_man\_pl\_waste\_kgPerday 3.650e-01 2.405e-17 1.518e+16 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.058e-10 on 187 degrees of freedom  
## Multiple R-squared: 1, Adjusted R-squared: 1   
## F-statistic: 1.507e+32 on 4 and 187 DF, p-value: < 2.2e-16

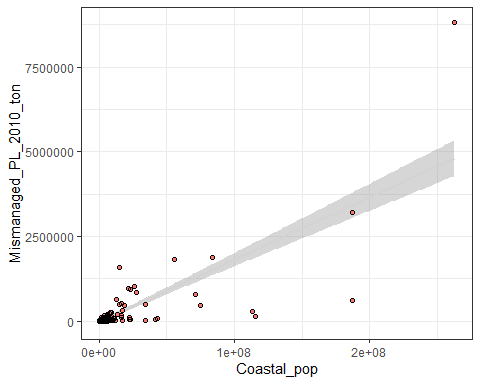
## statistic on correlation between total waste and plastic waste generation  
wasteSum <- sum(df$Waste\_gen\_kgPerday)  
plasticSum <- sum(df$Pl\_waste\_gen\_kgPerday)  
ratioPL\_W <- (plasticSum/wasteSum)\*100  
# 11.08% of the waste generated per kg per day is plastic. Percentage might not seem to high, but it seems  
# more than enaough yield very high correlation between waste generation and plastic generation.  
  
  
## plot1  
# here we can see the graphical representation of the plastic waste generation as a function of  
# waste generation. Lower waste generation shows high relation with plastic generation  
# which would imply that even limiting the waste generation, would still yield high plastic generation.  
p <- ggplot(dfN, aes(x = dfN$Waste\_gen\_kgPerday, y = dfN$Pl\_waste\_gen\_kgPerday))  
p <- p + geom\_smooth(method = "lm",  
 se = TRUE, color = "lightgrey")  
p <- p + geom\_segment(aes(xend = dfN$Waste\_gen\_kgPerday, yend = predicted),  
 alpha = .2) # alpha to fade lines  
p <- p + geom\_point(aes(color='red'),show.legend = FALSE)  
p <- p + geom\_point(aes(y = predicted), shape = 1)  
p <- p + xlab('Waste\_gen\_kgPerday') + ylab('Pl\_waste\_gen\_kgPerday')  
p <- p + theme\_bw()  
p



## plot2  
# Higher waste generation does not necesairilly mean high mismanage of plastic.  
p <- ggplot(dfN, aes(x = dfN$Waste\_gen\_kgPerday, y = dfN$Mismanaged\_PL\_2010\_ton))  
p <- p + geom\_smooth(method = "lm",  
 se = TRUE, color = "lightgrey")  
p <- p + geom\_segment(aes(xend = dfN$Waste\_gen\_kgPerday, yend = predicted),  
 alpha = .2) # alpha to fade lines  
p <- p + geom\_point(aes(color='red'),show.legend = FALSE)  
p <- p + geom\_point(aes(y = predicted), shape = 1)  
p <- p + xlab('Waste\_gen\_kgPerday') + ylab('Mismanaged\_PL\_2010\_ton')  
p <- p + theme\_bw()  
p



## plot3  
p <- ggplot(dfN, aes(x = dfN$Coastal\_pop, y = dfN$Mismanaged\_PL\_2010\_ton))  
p <- p + geom\_smooth(method = "lm",  
 se = TRUE, color = "lightgrey")  
p <- p + geom\_segment(aes(xend = dfN$Coastal\_pop, yend = predicted),  
 alpha = .5) # alpha to fade lines  
p <- p + geom\_point(aes(color='red'),show.legend = FALSE)  
p <- p + geom\_point(aes(y = predicted), shape = 1)  
p <- p + xlab('Coastal\_pop') + ylab('Mismanaged\_PL\_2010\_ton')  
p <- p + theme\_bw()  
p

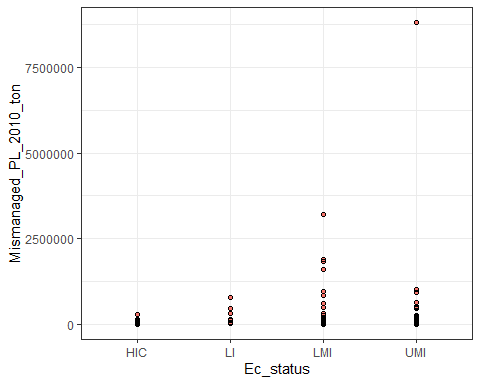


## lm-2  
dfN2 <- df[,-c(1,8)]  
myModel2=lm(dfN2$Mismanaged\_PL\_2010\_ton ~., data=dfN2)  
dfN2$predicted <- predict(myModel2)  
dfN2$residuals <- myModel2$residuals  
summary(myModel2)

## Warning in summary.lm(myModel2): essentially perfect fit: summary may be  
## unreliable

##   
## Call:  
## lm(formula = dfN2$Mismanaged\_PL\_2010\_ton ~ ., data = dfN2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.668e-09 -6.435e-11 2.530e-12 7.165e-11 1.998e-09   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.100e-10 3.308e-11 -3.327e+00 0.00106 \*\*   
## Ec\_statusLI 5.717e-10 6.734e-11 8.489e+00 6.84e-15 \*\*\*  
## Ec\_statusLMI 1.158e-10 5.313e-11 2.180e+00 0.03053 \*   
## Ec\_statusUMI -1.574e-10 4.944e-11 -3.184e+00 0.00171 \*\*   
## Coastal\_pop -7.803e-18 1.475e-18 -5.288e+00 3.47e-07 \*\*\*  
## Waste\_gen\_kgPerday 7.838e-18 3.173e-18 2.470e+00 0.01441 \*   
## Pl\_waste\_gen\_kgPerday 7.300e-03 2.400e-17 3.042e+14 < 2e-16 \*\*\*  
## Inad\_man\_pl\_waste\_kgPerday 3.650e-01 1.610e-17 2.267e+16 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.672e-10 on 184 degrees of freedom  
## Multiple R-squared: 1, Adjusted R-squared: 1   
## F-statistic: 1.987e+32 on 7 and 184 DF, p-value: < 2.2e-16

# df$predicted <- NULL  
# df$residuals <- NULL  
  
## plot1  
# Low-medium income countries seem to contribute a lot in Mismanaged\_PL\_2010\_ton  
p <- ggplot(dfN2, aes(x = dfN2$Ec\_status, y = dfN2$Mismanaged\_PL\_2010\_ton))  
p <- p + geom\_smooth(method = "lm",  
 se = TRUE, color = "lightgrey")  
p <- p + geom\_segment(aes(xend = dfN2$Ec\_status, yend = predicted),  
 alpha = .2) # alpha to fade lines  
p <- p + geom\_point(aes(color='red'),show.legend = FALSE)  
p <- p + geom\_point(aes(y = predicted), shape = 1)  
p <- p + xlab('Ec\_status') + ylab('Mismanaged\_PL\_2010\_ton')  
p <- p + theme\_bw()  
p



### SUM UP ###

# there seems to be clear evidence on income and high plastic pollution.  
# coastal population seems to play a role but it is not the most important one.  
# plastic consists of ~11% of total waste production. Seems low, but if one thinks   
# of the total production of waste per kg per day, 11% is still quite big a number.  
# lm and correlation suggest that plastic production is highly dependent on total waste production,   
# more than to other factors.  
# We could easily divide the dataset to train and test set. redevelop the models and apply them to the test set.  
# China clearly shows to be the catalyst in all of the research above, even when it acts out as an outlier.  
# I do not have time to do it today, but the lm developing above shows that it linear regression most probably suits our needs.  
# Further models could be applied out of the sphere of linear regression. I will not go over it  
# now cause we need something as solid as possible for tomorrow.  
# Models could also be developed to show the correlation between the factors and the   
# mismanaged waste of 2025 show as to produce our own model for what the authors have predicted,  
# though I don't know if such a thing would be worthwile.