**Part 1 – Setting goal and Finding data**

A study on Life Expectancy among countries from 2000-2015

Categories of factors that can affect life expectancy:

\* Social factors: Status, GDP, Income Composition of Resource, Schooling  
\* Mortality: Adult mortality, Infant death and Number of under-five deaths  
\* Healthcare-related: Immunization coverage (Polio, Hepatitis B, Measles and Diphtheria), Alcohol consumption, HIV/AIDS recorded cases, BMI and prevalence of thinness among children and adolescents.

**Objective:**

*To study on factors that are influencing life expectancy*

**Goal:**  
 *To conduct statistical analysis on life expectancy dataset by utilizing Regression and Classification model*

**Few Questions following regression will answer:**  
*1) Is there correlation between life expectancy and healthcare expenditures spent by each country?*  
*2) Is this country classified as a developing or developed country based on Life Expectancy value?*

**Structure and Summary of the Dataset**

#Load all required libraries for cleaning and analysis

library(dplyr)

library(stats)

library(e1071)

library(ggplot2)

library(forecast)

library(psych)

library(tidyr)

library(ExcelFunctionsR)

library(aod)

library(tidyverse)

library(modelr)

library(broom)

#Data reading

data <- read.csv ('./Life Expectancy Data.csv');

#Compactly Display the Structure of an Arbitrary R Object

str(data);

dim(data); #Dimensions of an object

summary(data) #Summary of the given data

**Part 2 - Collection and Cleaning Data**

# Make sure all the countries have data for all years (2000-2015)

# Count number of country

unique(data$Country) # will yield 193

unique(data$Year) # will yield 16

# Normally, with 193 countries and 16 years, we should have

193 \* 16 ## 3088 rows, but instead we have

length(data$Country) #2938 rows

countrycount <- data.frame(table(data$Country)) #countries without 16 counts will be excluded from the data set

# cleaning out N/A

is.na(data)

sum(is.na(data))

prelimlifexpect <- na.omit(data)

#filter final data set to be processed

flifexpect <- left\_join(prelimlifexpect, countrycount, by = c("Country" = "Var1"))

sum(is.na(flifexpect))

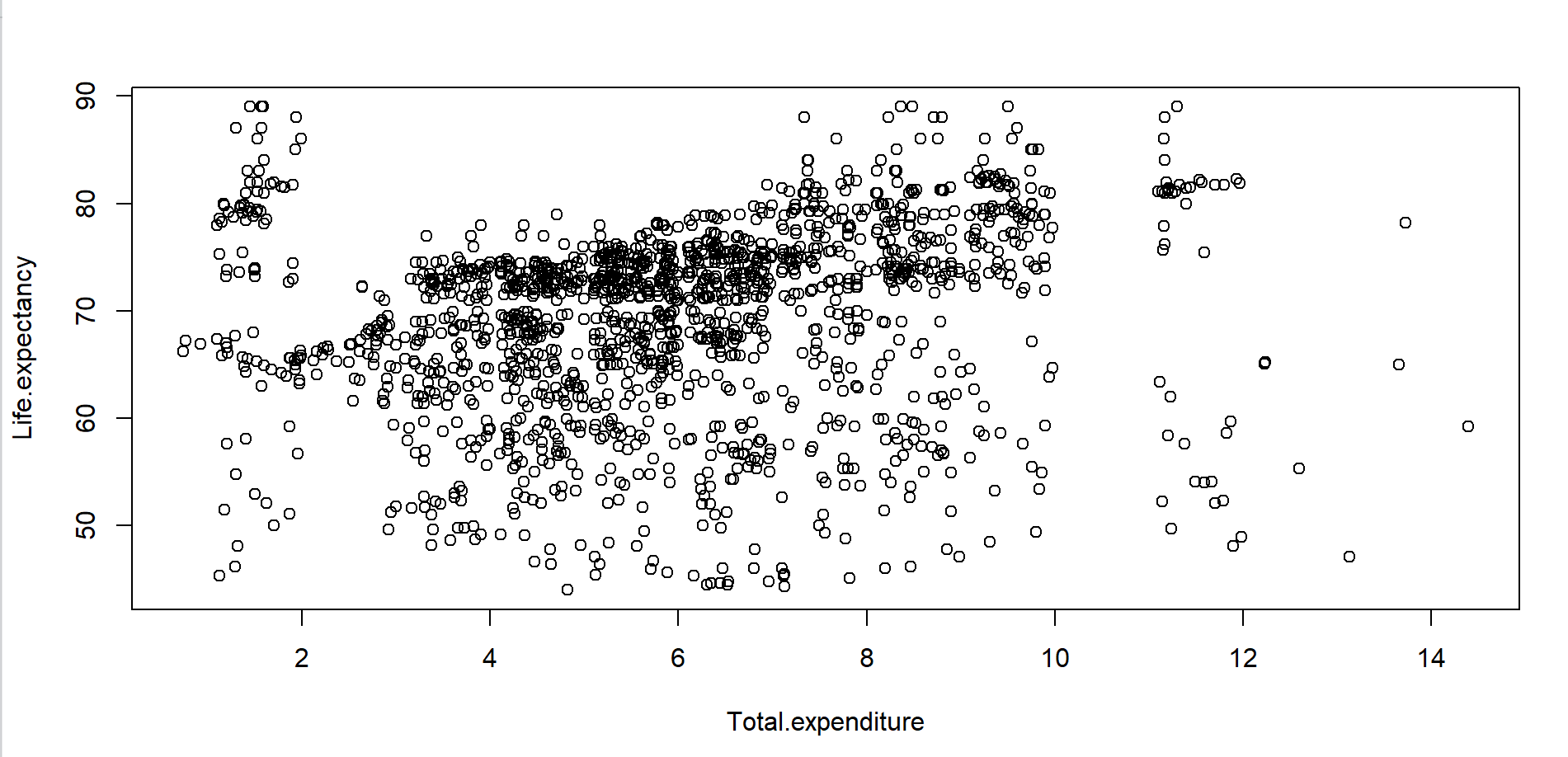
head(flifexpect)

**Part 3 – Regression Analysis**

# Correlation expenditure vs LifeExpectancy

par(mar=c(2,2,1,1))

plot(flifexpect$Total.expenditure, flifexpect$Life.expectancy)



corr.test(flifexpect$Total.expenditure, flifexpect$Life.expectancy)

corr.test(data$Total.expenditure, data$Life.expectancy)

# Comments: the correlation of 0.22 implies there is no strong correlation between healthcare expenditure and life expectancy

# in other words, life expectancy doesn't depend much on how much is spent on healthcare.

**Linear Regression**

y <- flifexpect$Life.expectancy

x <- flifexpect$Total.expenditure

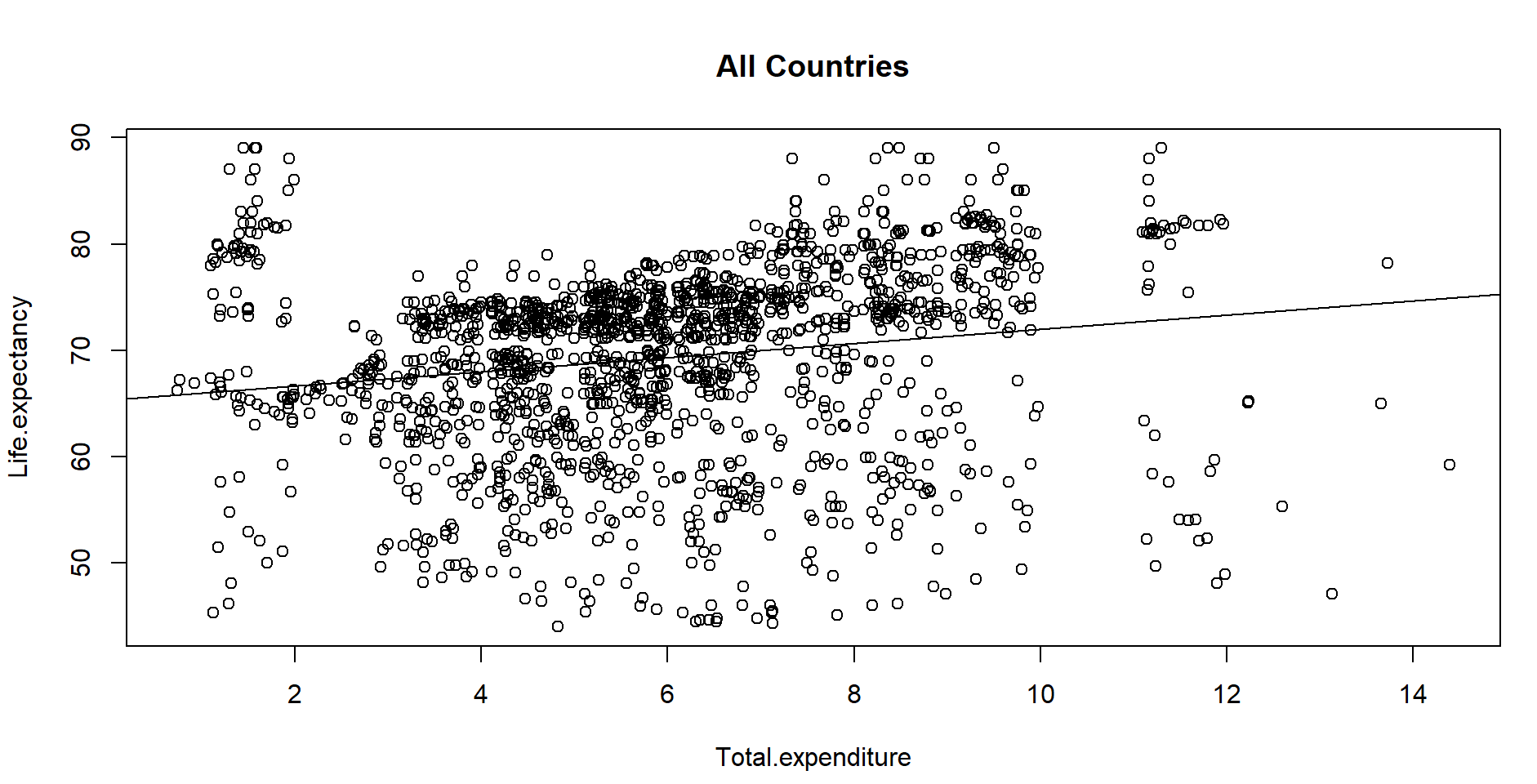
rmod <- lm(y ~ x)

summary(rmod)

attributes(rmod)

median(flifexpect$Life.expectancy, na.rm = TRUE)

median(flifexpect$Total.expenditure, na.rm = TRUE)



# From this plot, we can see that as the expenditure gets above the median, so does the life expectancy.

# There might be some skewness in the data that prohibit the establishment of a strong trend.

# We will classify data based on country status: Developing and Developed

**Classification**

# Determining the country status by looking at life expectancy

age\_group <- group\_by(flifexpect, flifexpect$Life.expectancy, na.rm = TRUE)

summarise(age\_group,

avg = mean(flifexpect$Status),

median = median(flifexpect$Status),

n = n())

write.csv(age\_group, "age\_group.csv")

# The average life expectancy (AVERAGEIF in Excel) gives the following results

# For Developing countries, the average life expectancy is 66.30 years

# For Developed countries, the average life expectancy is 79.65 years.

**Developing Countries**

dvping <- filter(age\_group, Status == "Developing")

ydeving <- dvping$Life.expectancy

xdeving <- dvping$Total.expenditure

rmoddeving <- lm(ydeving ~ xdeving)

summary(rmoddeving)

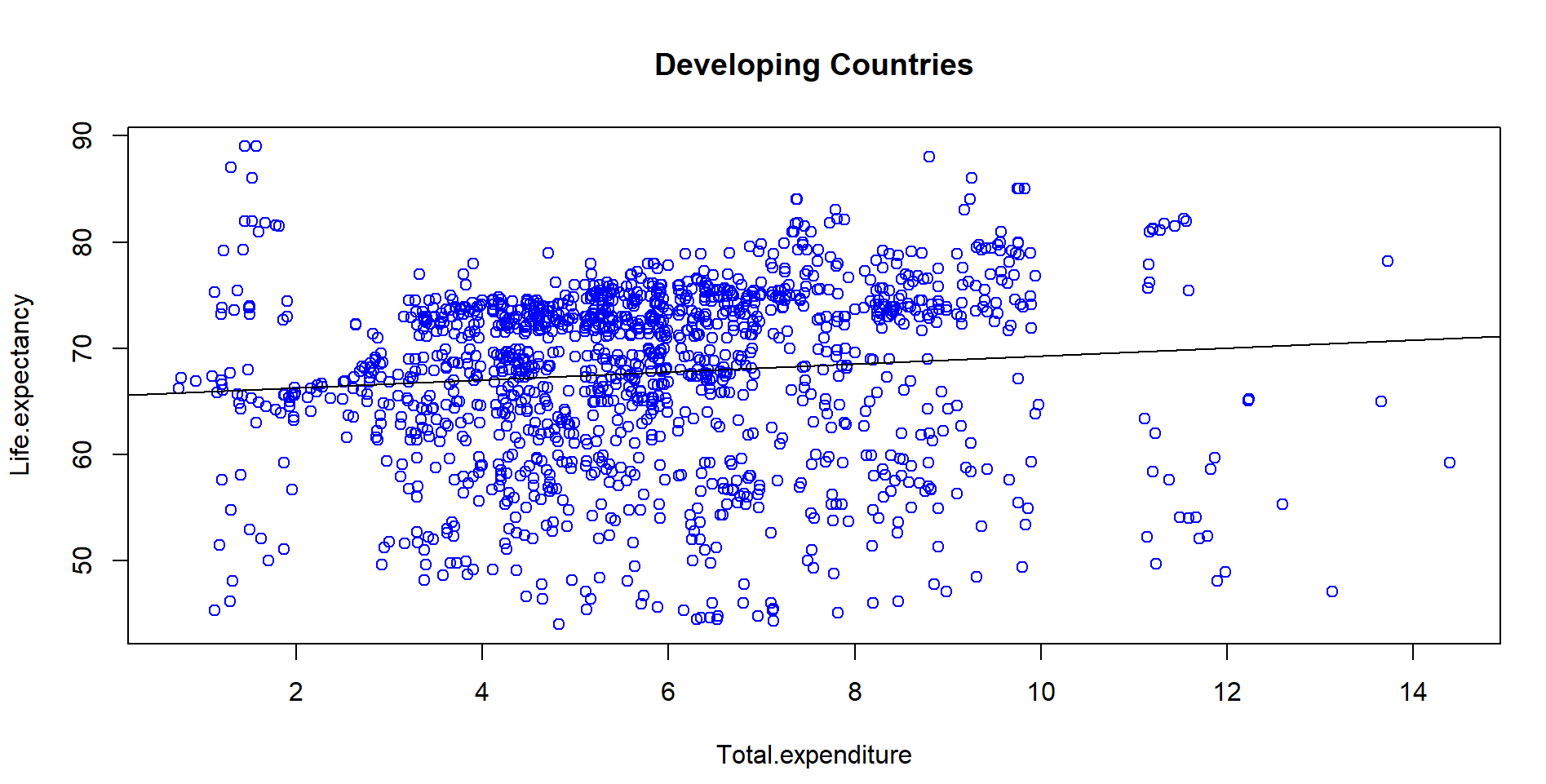
mean(dvping$Total.expenditure, na.rm = TRUE)

mean(dvping$Life.expectancy, na.rm = TRUE)

corr.test(dvping$Total.expenditure, dvping$Life.expectancy)

median(dvping$Total.expenditure, na.rm = TRUE)

median(dvping$Life.expectancy, na.rm = TRUE)



# Comments: At 4.92%, the p-value is still below 5% but it's way higher than the p-value for all countries data set.

# Hence still falling to confirm a dependence between total expenditure and life expectancy

# However, we observe that as the total expenditure cross its median value so does the life expectancy

#Developed Countries

dvped <- filter(age\_group, Status == "Developed")

ydvped <- dvped$Life.expectancy

xdvped <- dvped$Total.expenditure

rmoddvped <- lm(ydvped ~ xdvped)

summary(rmoddvped)

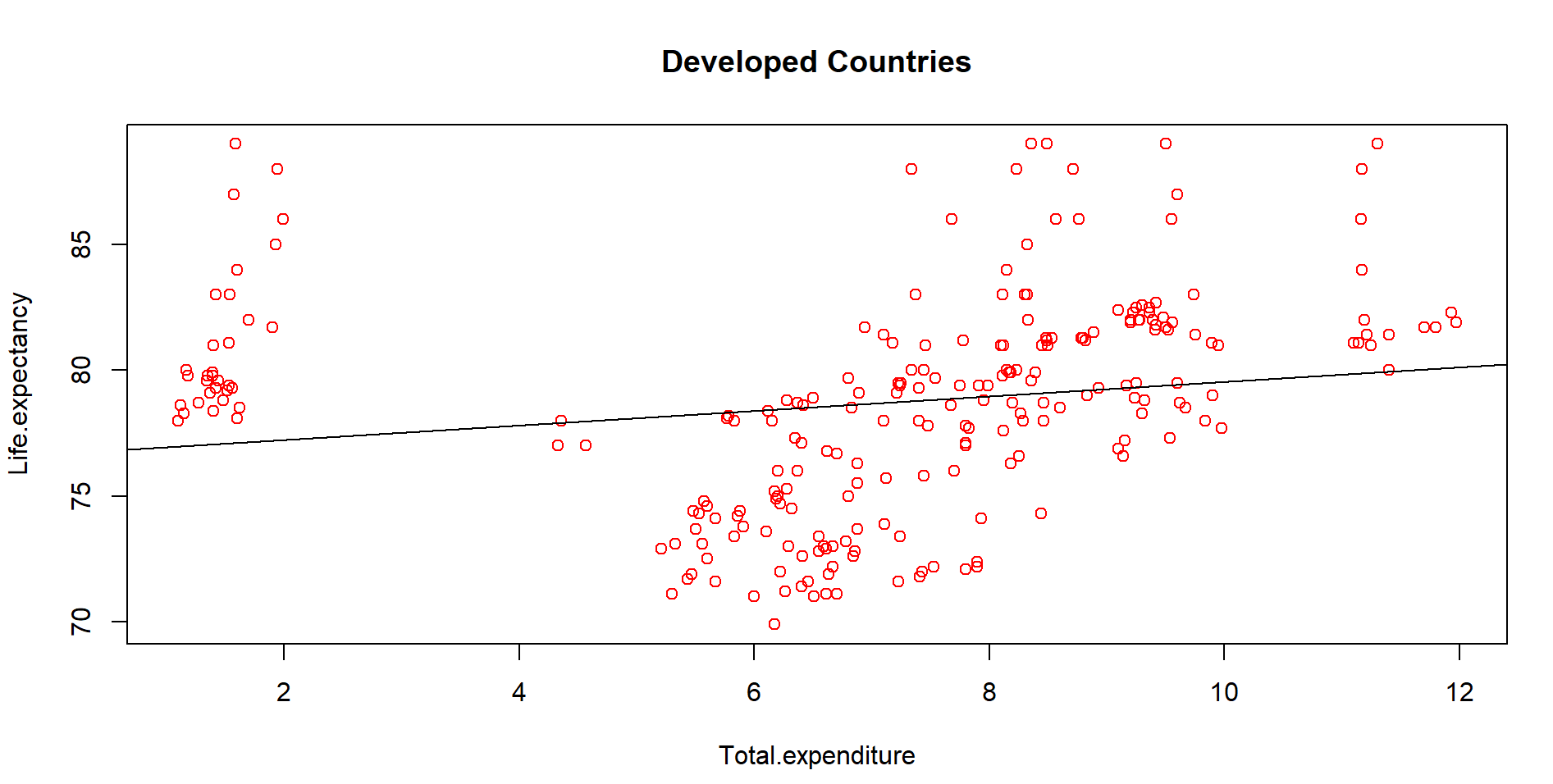
mean(dvped$Total.expenditure, na.rm = TRUE)

mean(dvped$Life.expectancy, na.rm = TRUE)

corr.test(dvped$Total.expenditure, dvped$Life.expectancy)

median(dvped$Total.expenditure, na.rm = TRUE)

median(dvped$Life.expectancy, na.rm = TRUE)



# Comments: Developed countries spend more than developing countries; developed countries also have a higher life expectancy.

**Part 4 - Classification Analysis**

#Brief summary of the cleaned data frame

head(flifexpect)

#Binary response (outcome, dependent) variable is Status

#Status - Developed or Developing status

#Predictor variable is Life Expectancy, GDP, Income Composition, Total Expenditure

#Life Expectancy - Life expectancy in Age

#Income Composition - Human Development Index in terms of income composition of resources (index ranging from 0 to 1)

#Total Expenditure - General government expenditure on health as a percentage of total government expenditure

#Prepare the data frame for the model

#Here I am taking the data for 2014 year

logdf <- flifexpect[flifexpect$Year == 2014, ]

row.names(logdf) <- logdf$Country

colnames(logdf)

#Subset the data frame for the classification model

logdf <- subset(logdf,select = c(Status, Life.expectancy,Income.composition.of.resources,Total.expenditure))

glimpse(logdf)

#Convert the status to binary

lookup <- c("Developing"=0, "Developed"=1)

logdf$binStatus <- lookup[logdf$Status]

glimpse(logdf)

#Drop the string Status column from data frame

logdf <- subset(logdf, select = -c(Status))

#Rename the binary status column to Status

names(logdf)[names(logdf) == "binStatus"] <- "Status"

glimpse(logdf)

#Rank each country by income composition of resources

#Wikipedia on HDI

#A value above 0.800 is classified as very high,

#between 0.700 and 0.799 high,

#0.550 to 0.699 as medium

#and anything below 0.550 as low

#Group each Income Composition into Rank Range

logdft <- logdf %>% mutate(HDI.Rank = cut(logdf$Income.composition.of.resources, c(0, 0.550, 0.700, 0.800, Inf)))

#Convert the HDI Rank to rank 1-4

# 1 - 0.800 and above - Very High

# 2 - 0.700 to 0.799 - High

# 3 - 0.550 to 0.699 - Medium

# 4 - 0.550 and below - Low

lookup2 <- c("(0,0.55]"=4, "(0.55,0.7]"=3, "(0.7,0.8]"=2, "(0.8,Inf]"=1)

logdft$ranked.HDI <- lookup2[logdft$HDI.Rank]

#Drop Income Composition and the HDI Rank column from data frame as it represented by ranked.HDI

logdft <- subset(logdft, select = -c(Income.composition.of.resources,HDI.Rank))

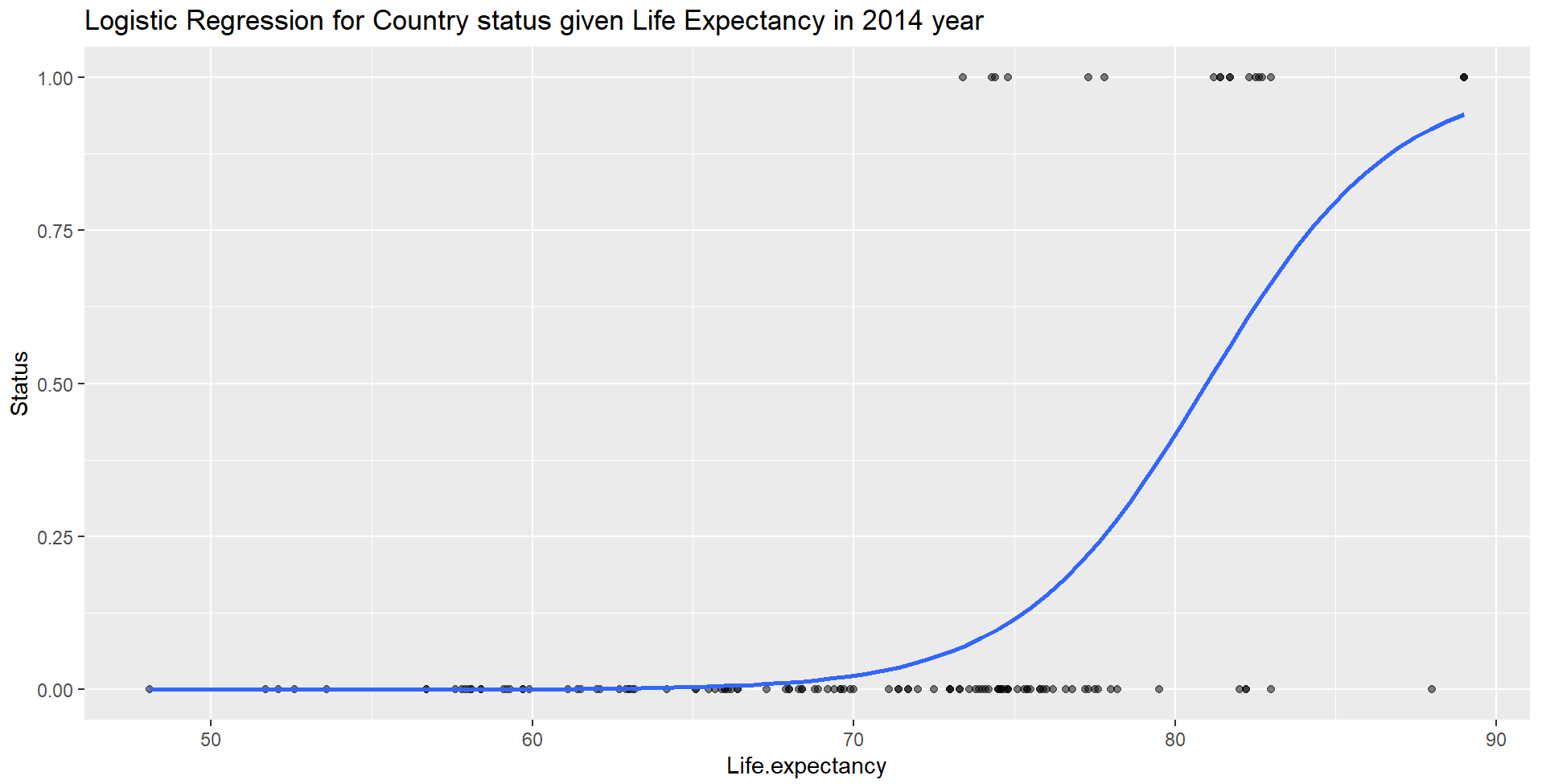
glimpse(logdft)

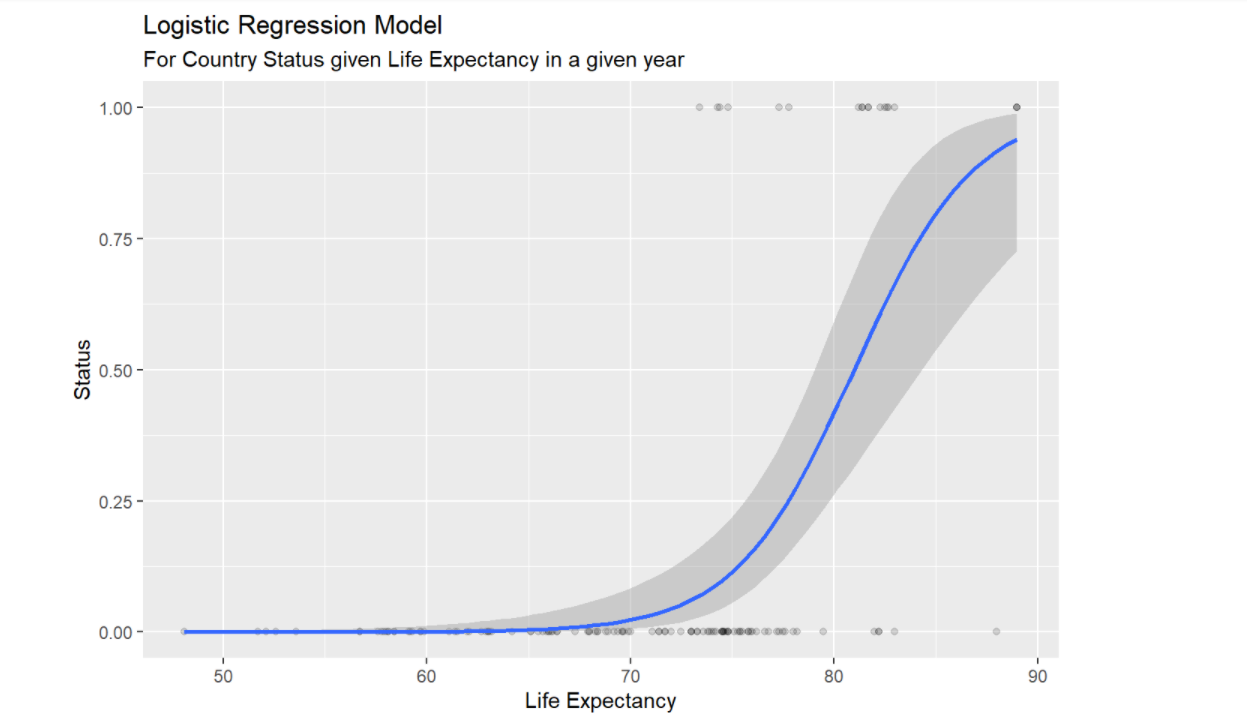
#Show summary of the data frame to ensure its good for logit model

summary(logdft)

#To predict the outcome of Status based on Life Expectancy value

# Logistics Regression





# Split the data into two chunks; training and testing set

# Regression is done on training set

train <- logdft[1:65,]

test <- logdft[66:131,]

# Simple Logistic Regression

model1 <- glm(Status ~ Life.expectancy, family = "binomial", data = train)

# Summary of the model

summary(model1)

# Accessing coefficients shows the coefficient estimates and related information that result from fitting a logistic regression model in order to predict the probability of Developed Status = Yes using Life Expectancy

tidy(model1)

# Interpretation of the balance coefficient for every increasing of life expectancy value by 0.1. The odds of the Status of a country increases by a factor of 1.3577

exp(coef(model1))

# Making prediction from the coefficients calculated, to compute the probability of Country's Status for any given Life Expectancy value. For example, given Life Expectancy are 68.9 and 73.7

predict(model1, data.frame(Life.expectancy = c(68.9, 73.7)), type = "response")

# From the result, for Country with Life Expectancy of 68.9, it is ~13.73% probability that it is a developed country. However, for Country with Life Expectancy of 73.7, it is more than 56.9% likely is a developed country, which is a high jump from mere 68.9 to 73.7