DS6372 Project 1

Caleb Thornsbury, Stephanie Duarte, Steven Cox

2024-02-02

## Life Expectancy Data Set

# Custom function to clean up any issues with column names  
processColumnNames <- function(data\_frame) {  
 # Apply the custom name processing logic to each column name  
 modifiedNames <- sapply(names(data\_frame), function(name) {  
 name <- tolower(name) # Lowercase all characters  
 name <- trimws(name) # Strip leading and trailing spaces  
 name <- gsub(" +", " ", name) # Replace multiple spaces with a single space  
 name <- gsub(" ", ".", name) # Replace single space with a period  
 return(name) # Return the modified name  
 })  
   
 # Ensure the modified column names are valid R identifiers and unique  
 uniqueNames <- make.names(modifiedNames, unique = TRUE)  
 names(data\_frame) <- uniqueNames # Update the data frame's column names  
   
 return(data\_frame)  
}  
  
# Import the CSV file without altering the column names  
life\_data <- read.csv("Data/Life Expectancy Data.csv", check.names = FALSE)  
# Clean up any column naming issues  
life\_data <- processColumnNames(life\_data)  
head(life\_data)

## country year status life.expectancy adult.mortality infant.deaths  
## 1 Afghanistan 2015 Developing 65.0 263 62  
## 2 Afghanistan 2014 Developing 59.9 271 64  
## 3 Afghanistan 2013 Developing 59.9 268 66  
## 4 Afghanistan 2012 Developing 59.5 272 69  
## 5 Afghanistan 2011 Developing 59.2 275 71  
## 6 Afghanistan 2010 Developing 58.8 279 74  
## alcohol percentage.expenditure hepatitis.b measles bmi under.five.deaths  
## 1 0.01 71.279624 65 1154 19.1 83  
## 2 0.01 73.523582 62 492 18.6 86  
## 3 0.01 73.219243 64 430 18.1 89  
## 4 0.01 78.184215 67 2787 17.6 93  
## 5 0.01 7.097109 68 3013 17.2 97  
## 6 0.01 79.679367 66 1989 16.7 102  
## polio total.expenditure diphtheria hiv.aids gdp population  
## 1 6 8.16 65 0.1 584.25921 33736494  
## 2 58 8.18 62 0.1 612.69651 327582  
## 3 62 8.13 64 0.1 631.74498 31731688  
## 4 67 8.52 67 0.1 669.95900 3696958  
## 5 68 7.87 68 0.1 63.53723 2978599  
## 6 66 9.20 66 0.1 553.32894 2883167  
## thinness.1.19.years thinness.5.9.years income.composition.of.resources  
## 1 17.2 17.3 0.479  
## 2 17.5 17.5 0.476  
## 3 17.7 17.7 0.470  
## 4 17.9 18.0 0.463  
## 5 18.2 18.2 0.454  
## 6 18.4 18.4 0.448  
## schooling  
## 1 10.1  
## 2 10.0  
## 3 9.9  
## 4 9.8  
## 5 9.5  
## 6 9.2

## Preliminary Data Wrangling

Explanatory variable is Life Expectancy. Therefore, before starting the EDA, any rows missing the explanatory variable will be filtered out.

missing <- life\_data %>% select(life.expectancy) %>% sapply(function(x) sum(is.na(x)))  
print(paste("Number of observations missing explanatory variables:", missing))

## [1] "Number of observations missing explanatory variables: 10"

#Remove all rows that have missing values of our explanatory variable life.expectancy  
life <- life\_data %>% filter(!is.na(life.expectancy))  
# To see the first few rows of the filtered data frame  
summary(life)

## country year status life.expectancy  
## Length:2928 Min. :2000 Length:2928 Min. :36.30   
## Class :character 1st Qu.:2004 Class :character 1st Qu.:63.10   
## Mode :character Median :2008 Mode :character Median :72.10   
## Mean :2008 Mean :69.22   
## 3rd Qu.:2011 3rd Qu.:75.70   
## Max. :2015 Max. :89.00   
##   
## adult.mortality infant.deaths alcohol percentage.expenditure  
## Min. : 1.0 Min. : 0.00 Min. : 0.010 Min. : 0.000   
## 1st Qu.: 74.0 1st Qu.: 0.00 1st Qu.: 0.905 1st Qu.: 4.854   
## Median :144.0 Median : 3.00 Median : 3.770 Median : 65.611   
## Mean :164.8 Mean : 30.41 Mean : 4.615 Mean : 740.321   
## 3rd Qu.:228.0 3rd Qu.: 22.00 3rd Qu.: 7.715 3rd Qu.: 442.614   
## Max. :723.0 Max. :1800.00 Max. :17.870 Max. :19479.912   
## NA's :193   
## hepatitis.b measles bmi under.five.deaths  
## Min. : 1.00 Min. : 0.0 Min. : 1.00 Min. : 0.00   
## 1st Qu.:77.00 1st Qu.: 0.0 1st Qu.:19.30 1st Qu.: 0.00   
## Median :92.00 Median : 17.0 Median :43.35 Median : 4.00   
## Mean :80.96 Mean : 2427.9 Mean :38.24 Mean : 42.18   
## 3rd Qu.:97.00 3rd Qu.: 362.2 3rd Qu.:56.10 3rd Qu.: 28.00   
## Max. :99.00 Max. :212183.0 Max. :77.60 Max. :2500.00   
## NA's :553 NA's :32   
## polio total.expenditure diphtheria hiv.aids   
## Min. : 3.00 Min. : 0.37 Min. : 2.00 Min. : 0.100   
## 1st Qu.:78.00 1st Qu.: 4.26 1st Qu.:78.00 1st Qu.: 0.100   
## Median :93.00 Median : 5.75 Median :93.00 Median : 0.100   
## Mean :82.55 Mean : 5.93 Mean :82.32 Mean : 1.748   
## 3rd Qu.:97.00 3rd Qu.: 7.49 3rd Qu.:97.00 3rd Qu.: 0.800   
## Max. :99.00 Max. :17.60 Max. :99.00 Max. :50.600   
## NA's :19 NA's :226 NA's :19   
## gdp population thinness.1.19.years thinness.5.9.years  
## Min. : 1.68 Min. :3.400e+01 Min. : 0.100 Min. : 0.100   
## 1st Qu.: 463.85 1st Qu.:1.967e+05 1st Qu.: 1.600 1st Qu.: 1.575   
## Median : 1764.97 Median :1.392e+06 Median : 3.300 Median : 3.400   
## Mean : 7494.21 Mean :1.276e+07 Mean : 4.851 Mean : 4.881   
## 3rd Qu.: 5932.90 3rd Qu.:7.427e+06 3rd Qu.: 7.200 3rd Qu.: 7.200   
## Max. :119172.74 Max. :1.294e+09 Max. :27.700 Max. :28.600   
## NA's :443 NA's :644 NA's :32 NA's :32   
## income.composition.of.resources schooling   
## Min. :0.0000 Min. : 0.0   
## 1st Qu.:0.4930 1st Qu.:10.1   
## Median :0.6770 Median :12.3   
## Mean :0.6274 Mean :12.0   
## 3rd Qu.:0.7792 3rd Qu.:14.3   
## Max. :0.9480 Max. :20.7   
## NA's :160 NA's :160

source("Misc/QQvsLogQQ.r")

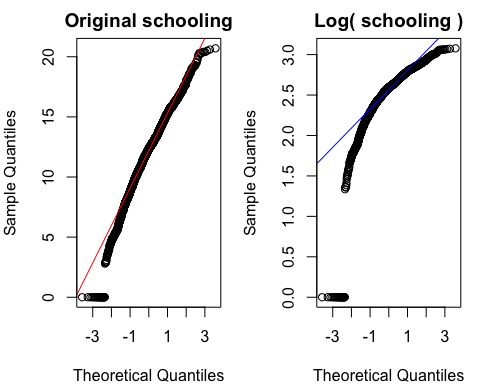
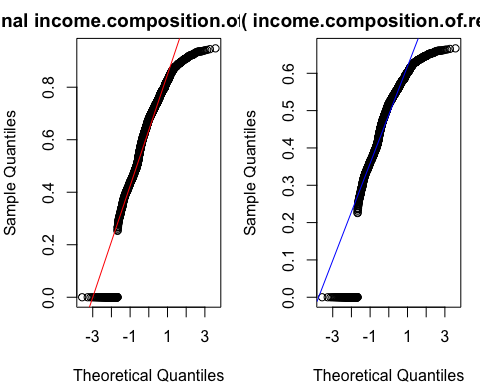
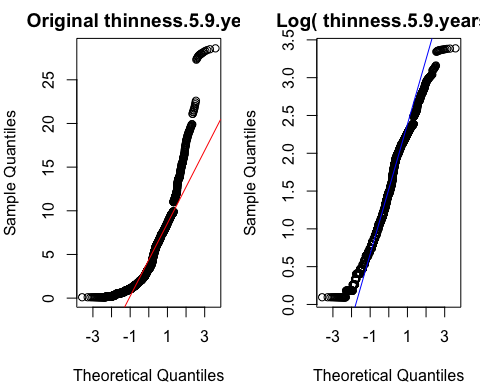
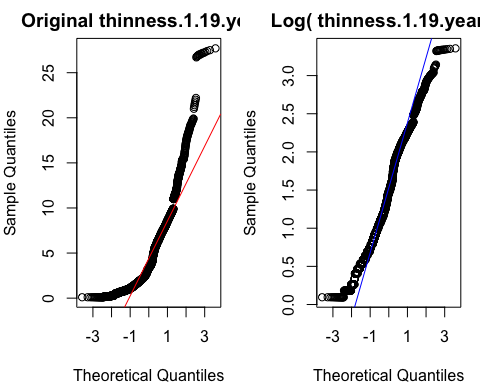
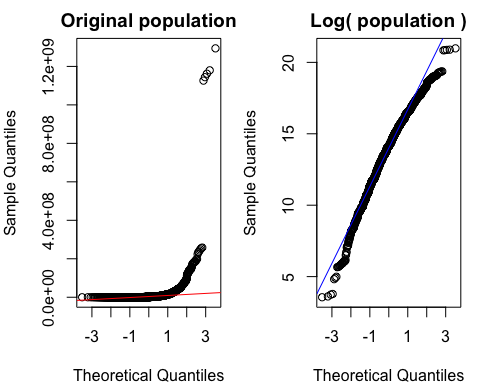
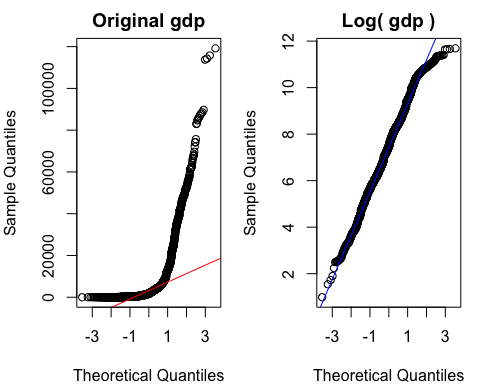
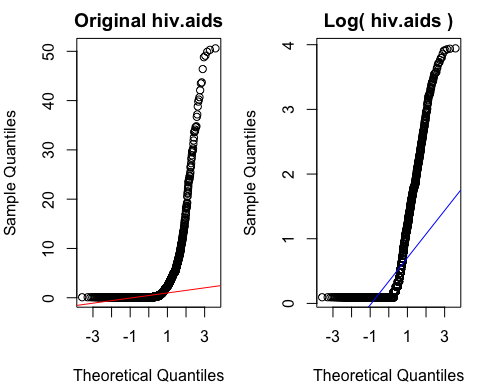
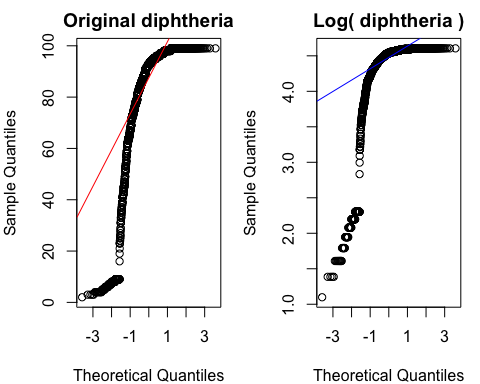
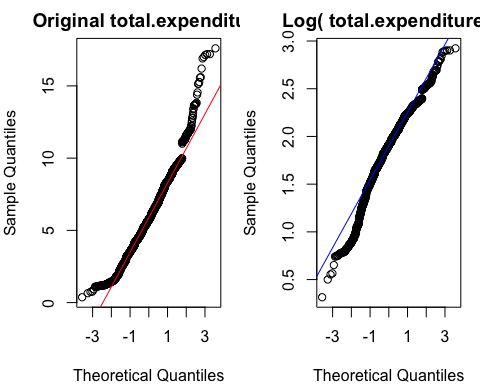
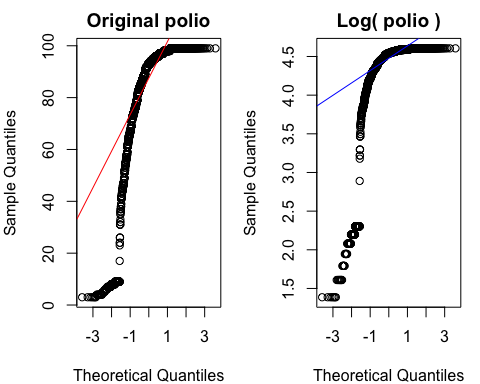
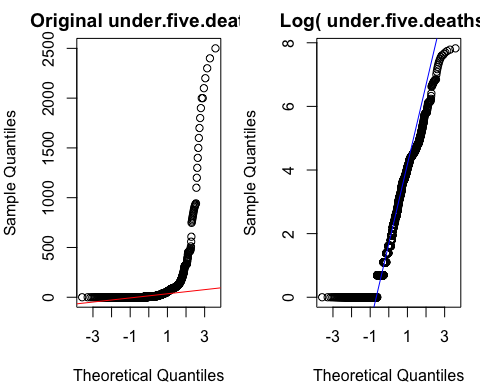
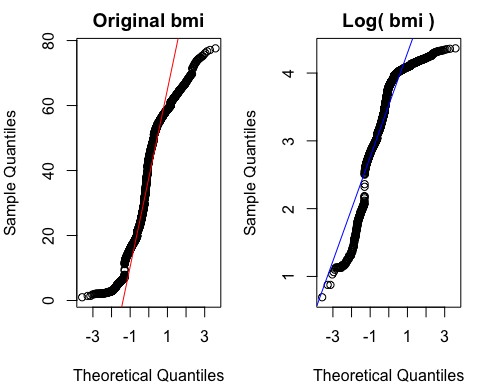
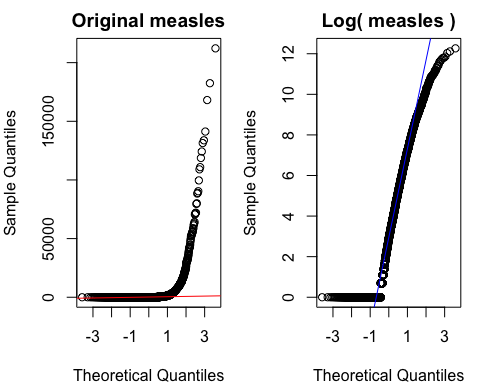
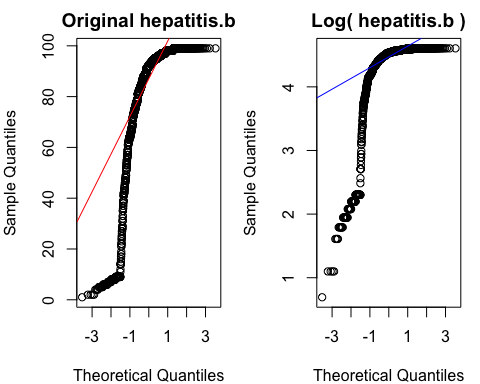
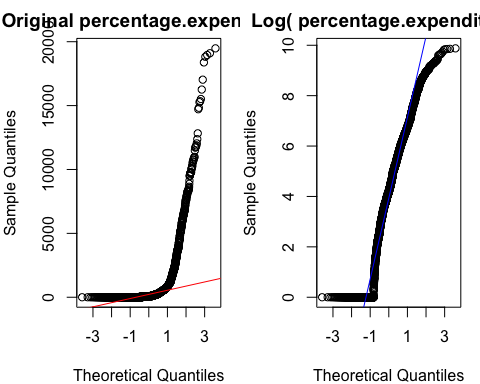
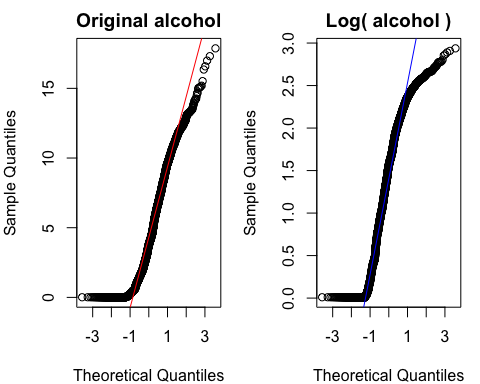
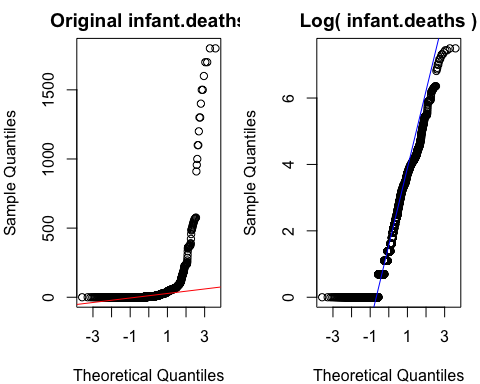
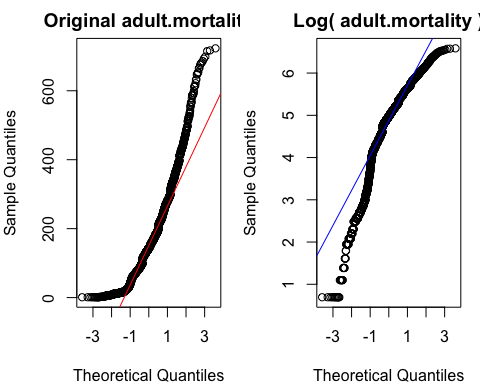
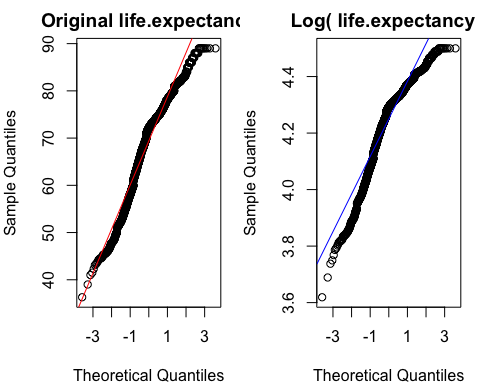
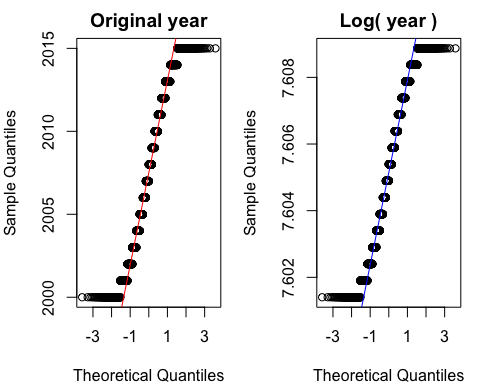
##   
## Attaching package: 'tidyr'

## The following objects are masked from 'package:Matrix':  
##   
## expand, pack, unpack

##   
## Attaching package: 'purrr'

## The following object is masked from 'package:car':  
##   
## some

## The following object is masked from 'package:caret':  
##   
## lift



Section to describe the data set and explanatory variable.

# Check uniqueness of each column  
life %>% summarise(across(everything(), ~n\_distinct(.)))

## country year status life.expectancy adult.mortality infant.deaths alcohol  
## 1 183 16 2 362 425 209 1077  
## percentage.expenditure hepatitis.b measles bmi under.five.deaths polio  
## 1 2323 88 958 603 252 74  
## total.expenditure diphtheria hiv.aids gdp population thinness.1.19.years  
## 1 817 82 200 2486 2278 201  
## thinness.5.9.years income.composition.of.resources schooling  
## 1 208 626 174

summary(life)

## country year status life.expectancy  
## Length:2928 Min. :2000 Length:2928 Min. :36.30   
## Class :character 1st Qu.:2004 Class :character 1st Qu.:63.10   
## Mode :character Median :2008 Mode :character Median :72.10   
## Mean :2008 Mean :69.22   
## 3rd Qu.:2011 3rd Qu.:75.70   
## Max. :2015 Max. :89.00   
##   
## adult.mortality infant.deaths alcohol percentage.expenditure  
## Min. : 1.0 Min. : 0.00 Min. : 0.010 Min. : 0.000   
## 1st Qu.: 74.0 1st Qu.: 0.00 1st Qu.: 0.905 1st Qu.: 4.854   
## Median :144.0 Median : 3.00 Median : 3.770 Median : 65.611   
## Mean :164.8 Mean : 30.41 Mean : 4.615 Mean : 740.321   
## 3rd Qu.:228.0 3rd Qu.: 22.00 3rd Qu.: 7.715 3rd Qu.: 442.614   
## Max. :723.0 Max. :1800.00 Max. :17.870 Max. :19479.912   
## NA's :193   
## hepatitis.b measles bmi under.five.deaths  
## Min. : 1.00 Min. : 0.0 Min. : 1.00 Min. : 0.00   
## 1st Qu.:77.00 1st Qu.: 0.0 1st Qu.:19.30 1st Qu.: 0.00   
## Median :92.00 Median : 17.0 Median :43.35 Median : 4.00   
## Mean :80.96 Mean : 2427.9 Mean :38.24 Mean : 42.18   
## 3rd Qu.:97.00 3rd Qu.: 362.2 3rd Qu.:56.10 3rd Qu.: 28.00   
## Max. :99.00 Max. :212183.0 Max. :77.60 Max. :2500.00   
## NA's :553 NA's :32   
## polio total.expenditure diphtheria hiv.aids   
## Min. : 3.00 Min. : 0.37 Min. : 2.00 Min. : 0.100   
## 1st Qu.:78.00 1st Qu.: 4.26 1st Qu.:78.00 1st Qu.: 0.100   
## Median :93.00 Median : 5.75 Median :93.00 Median : 0.100   
## Mean :82.55 Mean : 5.93 Mean :82.32 Mean : 1.748   
## 3rd Qu.:97.00 3rd Qu.: 7.49 3rd Qu.:97.00 3rd Qu.: 0.800   
## Max. :99.00 Max. :17.60 Max. :99.00 Max. :50.600   
## NA's :19 NA's :226 NA's :19   
## gdp population thinness.1.19.years thinness.5.9.years  
## Min. : 1.68 Min. :3.400e+01 Min. : 0.100 Min. : 0.100   
## 1st Qu.: 463.85 1st Qu.:1.967e+05 1st Qu.: 1.600 1st Qu.: 1.575   
## Median : 1764.97 Median :1.392e+06 Median : 3.300 Median : 3.400   
## Mean : 7494.21 Mean :1.276e+07 Mean : 4.851 Mean : 4.881   
## 3rd Qu.: 5932.90 3rd Qu.:7.427e+06 3rd Qu.: 7.200 3rd Qu.: 7.200   
## Max. :119172.74 Max. :1.294e+09 Max. :27.700 Max. :28.600   
## NA's :443 NA's :644 NA's :32 NA's :32   
## income.composition.of.resources schooling   
## Min. :0.0000 Min. : 0.0   
## 1st Qu.:0.4930 1st Qu.:10.1   
## Median :0.6770 Median :12.3   
## Mean :0.6274 Mean :12.0   
## 3rd Qu.:0.7792 3rd Qu.:14.3   
## Max. :0.9480 Max. :20.7   
## NA's :160 NA's :160