Data Cleaning and Analysis of Cervical Cancer Data

Advaith Cheruvu

9/21/2021

Importance of Study

There will be an estimated 14,480 new cases of invasive cervical cancer will be diagnosed and about 4,290 of these women will die from cervical cancer in 2021. This, along with the fact that cervical cancer is frequently diagnosed between the ages of 35-44, shows the importance of studying this disease. As patterns emerge from analyzing the data from susceptible women, we can adapt our health care system to better prevent and treat cervical cancer. For example, by knowing the risk factors, doctors can steer susceptible women away from the risk factors as a measure of prevention. Also, we can critique our current cancer diagnosis techniques so that we know which tests are the most effective and if the medical community needs to develop more accurate tests.

Overview of Cervical Cancer

Cervical cancer is a type of cancer that occurs in a woman's cervix. The main cause of cervical cancer is the human papilomavirus (HPV). Cervical cancer has 2 main types: squamous cell carcinoma and adenocarcinoma. Squamous cell carcinoma begins in the thin flat cells that line the outer part of the cervix. Most cases of cervical cancer are squamous cell carcinomas. Adenocarcinoma begins in the column-shaped glandular cells lining the cervical canal. Both of these forms of cervical cancer produce similar symtoms. Early stages generally have no signs or symptoms, but more advanced cervical cancers may include vaginal bleeding after intercourse, water/bloody vaginal discharge, and pelvic pain or pain during intercourse. Once these symptoms present themselves, the cancer would have already started spreading to nearby tissue and forming a tumor. Eventually, if left untreated, the cancer can spread to the bladder, intestines, bones, lymphnodes, lungs, and liver, leading to death.

HPV

HPV is the main cause of cervical cancer, but it may lie dormant for years before becoming active. Women may contract HPV from sexual intercourse, which may develop into cervical cancer. HPV is very common with at least half of sexually active people contracting it, but only relatively few women will get cervical cancer. In fact, most HPV cases clear within a few months without any intervention and 90% clear within 2 years. The other 10% of cases, however, have a high risk of progressing into cervical cancer. Preventing the transmission and damage done by this STD is crucial to preventing cervical cancer.

Prevention: Vaccination

There are currently 3 vaccines that protect against HPV 16 and HPV 18. This strands are known to cause at least 70% of cervical cancers. The third vaccine protects against 5 additional types of oncogenic HPV strands, which cause another 20% of cervical cancers. Also, two of the vaccines protect against HPV 6 and HPV 11, which are known to cause anogenital warts. These vaccines are best adminstered prior to the exposure of HPV, which is the reasoning behind vaccinating girls aged 9 to 14, before most have started sexual activity. These vaccines cannot treat an ongoing HPV infection or HPV-associated diseases. Though these vaccines are effective, it does not replace cervical cancer screenings.

Prevention: Screening

Cervical cancer screenings involve testing for pre-cancer and cancer in healthy women who have no symptoms. When screenings detect an HPV infection or pre-cancerous lesions, they can be easily treated and the risk of cancer can be avoided. Even screenings that detect cancer at an early stage has a high potential for effective treatment and cure. Screening is recomended for women aged 30 at regular intervals since HPV takes may take many years to progress.

Prevention: Abstinence

Since HPV is a sexually transmitted disease (STD), abstinence is the most effective prevention measure against HPV and cervical cancer. Limiting sexual partners and maintaining sexual activies in a long-term monogamous relationship reduces the risk of spreading HPV. Safe, protective sex can also reduce the risk of spreading HPV, but the other methods discussed earlier are more effective since HPV spreads very quickly during sexually intercourse.

Hinselmann Test

The Hinselmann test tests for cervical cancer by examining cells on colposcope. The colposcope gathers a sample of cells from the cervix and tests for cancerous growths. This test can diagnose cervical cancer and is done if the patient is showing symptoms or suspected to have cervical cancer based on the discretion of the personal care provider.

Schiller Test

The Schiller test tests for cervical cancer by applying iodine to the cervix. The iodine stains healthy cells brown, leaving abnormal cells unstained, usually appearing white or yellow. This test is mainly for larger, more noticeable cancer, allowing doctors to confirm the diagnosis of cervical cancer.

Cytosis

The cytology test includes a pap test or pap smear where the physician will gently remove cells from the cervix so that they may be checked under a microscope. Pap smears are the most common form of screening/diagnosing for cervical cancer due to the cost and ease of the procedure.

Biopsy

A cervical biopsy remove a sample of tissue for testing using a variety of techniques to obtain the sample. Punch biopsy uses a circular blade to obtain a hole-shaped piece of tissue, cone biopsy uses a laser or scalpel to remove a cone-shaped piece of tissue, and endocervical curettage (ECC) uses a curette to scape the inner lining of the endocervical canal.

Introduction to the Dataset

As mentioned earlier, cervical cancer is important to study; especially the data associated with susceptible women. The dataset I will be analyzing will be from the 8th Iberian Pattern Recognition and Image Analsis Conference. This dataset studied 858 women and some of the symptoms and behaviors of women that put them at risk cervical cancer or other similar diseases. I will go through the steps I took to clean and analyze the data.

Setting up Workspace and Installing Packages

The code below shows the preparation of the workspace and cleaning up the environment, as well as installing and loading packages. “tidyverse” is an important package that includes “dplyr” which helps with data science work. Some of the organizational tools and commands from “dplyr” will be shown later. Note that the “install.packages(”tidyverse“)” line is commented because this package has already been installed on this machine.

```
#
# Clean up and set up
rm(list=ls())
setwd("/Users/advai/Documents/DSFS")
#
# install and load libraries
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(stringr)
source("myfunctions.R")
```

Loading and Looking at Cervical Cancer Data

After reading the cervical cancer data into R studio, I can look at it using a variety of commands. The most important commands are “summary” which reports the quartile data, “str” which reports the type of data in the dataframe, and “names” which reports the names of the columns of data. Also, the “na.strings” argument replaces the “?” characters with “NA” so that they are easier to work with when I replace missing data.

```
#
# load cervical cancer data
cervicalCA <- read.csv(file = "C:\\Users\\advai\\Documents\\DSFS\\cervicalCA.csv",header=T, na.strings=c
("?"))
#
# looking at the data
names(cervicalCA)
```

```
##  [1] "Age"                                 "Number.of.sexual.partners"
##  [3] "First.sexual.intercourse"            "Num.of.pregnancies"
##  [5] "Smokes"                              "Smokes..years."
##  [7] "Smokes..packs.year."                 "Hormonal.Contraceptives"
##  [9] "Hormonal.Contraceptives..years."     "IUD"
## [11] "IUD..years."                         "STDs"
## [13] "STDs..number."                       "STDs.condylomatosis"
## [15] "STDs.cervical.condylomatosis"        "STDs.vaginal.condylomatosis"
## [17] "STDs.vulvo.perineal.condylomatosis" "STDs.syphilis"
## [19] "STDs.pelvic.inflammatory.disease"    "STDs.genital.herpes"
## [21] "STDs.molluscum.contagiosum"          "STDs.AIDS"
## [23] "STDs.HIV"                            "STDs.Hepatitis.B"
## [25] "STDs.HPV"                            "STDs..Number.of.diagnosis"
## [27] "STDs..Time.since.first.diagnosis"    "STDs..Time.since.last.diagnosis"
## [29] "Dx.Cancer"                           "Dx.CIN"
## [31] "Dx.HPV"                              "Dx"
## [33] "Hinselmann"                          "Schiller"
## [35] "Citology"                            "Biopsy"
```

```
summary(cervicalCA)
```

```
##       Age        Number.of.sexual.partners First.sexual.intercourse
##  Min.   :13.00   Min.   : 1.000            Min.   :10
##  1st Qu.:20.00   1st Qu.: 2.000            1st Qu.:15
##  Median :25.00   Median : 2.000            Median :17
##  Mean   :26.82   Mean   : 2.528            Mean   :17
##  3rd Qu.:32.00   3rd Qu.: 3.000            3rd Qu.:18
##  Max.   :84.00   Max.   :28.000            Max.   :32
##                  NA's   :26                NA's   :7
##  Num.of.pregnancies     Smokes       Smokes..years.  Smokes..packs.year.
##  Min.   : 0.000     Min.   :0.0000   Min.   : 0.00   Min.   : 0.0000
##  1st Qu.: 1.000     1st Qu.:0.0000   1st Qu.: 0.00   1st Qu.: 0.0000
##  Median : 2.000     Median :0.0000   Median : 0.00   Median : 0.0000
##  Mean   : 2.276     Mean   :0.1456   Mean   : 1.22   Mean   : 0.4531
##  3rd Qu.: 3.000     3rd Qu.:0.0000   3rd Qu.: 0.00   3rd Qu.: 0.0000
##  Max.   :11.000     Max.   :1.0000   Max.   :37.00   Max.   :37.0000
##  NA's   :56         NA's   :13       NA's   :13      NA's   :13
##  Hormonal.Contraceptives Hormonal.Contraceptives..years.      IUD
##  Min.   :0.0000          Min.   : 0.000                  Min.   :0.000
##  1st Qu.:0.0000          1st Qu.: 0.000                  1st Qu.:0.000
##  Median :1.0000          Median : 0.500                  Median :0.000
##  Mean   :0.6413          Mean   : 2.256                  Mean   :0.112
##  3rd Qu.:1.0000          3rd Qu.: 3.000                  3rd Qu.:0.000
##  Max.   :1.0000          Max.   :30.000                  Max.   :1.000
##  NA's   :108             NA's   :108                     NA's   :117
##   IUD..years.           STDs        STDs..number.    STDs.condylomatosis
##  Min.   : 0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.00000
##  1st Qu.: 0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.00000
##  Median : 0.0000   Median :0.0000   Median :0.0000   Median :0.00000
##  Mean   : 0.5148   Mean   :0.1049   Mean   :0.1766   Mean   :0.05843
##  3rd Qu.: 0.0000   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.00000
##  Max.   :19.0000   Max.   :1.0000   Max.   :4.0000   Max.   :1.00000
##  NA's   :117       NA's   :105      NA's   :105      NA's   :105
##  STDs.cervical.condylomatosis STDs.vaginal.condylomatosis
##  Min.   :0                    Min.   :0.00000
##  1st Qu.:0                    1st Qu.:0.00000
##  Median :0                    Median :0.00000
##  Mean   :0                    Mean   :0.00531
##  3rd Qu.:0                    3rd Qu.:0.00000
##  Max.   :0                    Max.   :1.00000
##  NA's   :105                  NA's   :105
##  STDs.vulvo.perineal.condylomatosis STDs.syphilis
##  Min.   :0.0000                     Min.   :0.0000
##  1st Qu.:0.0000                     1st Qu.:0.0000
##  Median :0.0000                     Median :0.0000
##  Mean   :0.0571                     Mean   :0.0239
##  3rd Qu.:0.0000                     3rd Qu.:0.0000
##  Max.   :1.0000                     Max.   :1.0000
##  NA's   :105                        NA's   :105
##  STDs.pelvic.inflammatory.disease STDs.genital.herpes
##  Min.   :0.00000                  Min.   :0.00000
##  1st Qu.:0.00000                  1st Qu.:0.00000
##  Median :0.00000                  Median :0.00000
##  Mean   :0.00133                  Mean   :0.00133
##  3rd Qu.:0.00000                  3rd Qu.:0.00000
##  Max.   :1.00000                  Max.   :1.00000
##  NA's   :105                      NA's   :105
##  STDs.molluscum.contagiosum   STDs.AIDS      STDs.HIV      STDs.Hepatitis.B
```

```
##  Min.   :0.00000            Min.   :0     Min.   :0.0000   Min.   :0.00000
##  1st Qu.:0.00000            1st Qu.:0     1st Qu.:0.0000   1st Qu.:0.00000
##  Median :0.00000            Median :0     Median :0.0000   Median :0.00000
##  Mean   :0.00133            Mean   :0     Mean   :0.0239   Mean   :0.00133
##  3rd Qu.:0.00000            3rd Qu.:0     3rd Qu.:0.0000   3rd Qu.:0.00000
##  Max.   :1.00000            Max.   :0     Max.   :1.0000   Max.   :1.00000
##  NA's   :105                NA's   :105   NA's   :105      NA's   :105
##     STDs.HPV       STDs..Number.of.diagnosis STDs..Time.since.first.diagnosis
##  Min.   :0.00000   Min.   :0.00000           Min.   : 1.000
##  1st Qu.:0.00000   1st Qu.:0.00000           1st Qu.: 2.000
##  Median :0.00000   Median :0.00000           Median : 4.000
##  Mean   :0.00266   Mean   :0.08741           Mean   : 6.141
##  3rd Qu.:0.00000   3rd Qu.:0.00000           3rd Qu.: 8.000
##  Max.   :1.00000   Max.   :3.00000           Max.   :22.000
##  NA's   :105                                 NA's   :787
##  STDs..Time.since.last.diagnosis   Dx.Cancer           Dx.CIN
##  Min.   : 1.000                  Min.   :0.00000   Min.   :0.00000
##  1st Qu.: 2.000                  1st Qu.:0.00000   1st Qu.:0.00000
##  Median : 3.000                  Median :0.00000   Median :0.00000
##  Mean   : 5.817                  Mean   :0.02098   Mean   :0.01049
##  3rd Qu.: 7.500                  3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :22.000                  Max.   :1.00000   Max.   :1.00000
##  NA's   :787
##      Dx.HPV              Dx            Hinselmann         Schiller
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000
##  Median :0.00000   Median :0.00000   Median :0.00000   Median :0.00000
##  Mean   :0.02098   Mean   :0.02797   Mean   :0.04079   Mean   :0.08625
##  3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.00000   Max.   :1.00000
##
##     Citology           Biopsy
##  Min.   :0.00000   Min.   :0.0000
##  1st Qu.:0.00000   1st Qu.:0.0000
##  Median :0.00000   Median :0.0000
##  Mean   :0.05128   Mean   :0.0641
##  3rd Qu.:0.00000   3rd Qu.:0.0000
##  Max.   :1.00000   Max.   :1.0000
##
```

```
str(cervicalCA)
```

```
## 'data.frame':    858 obs. of  36 variables:
##  $ Age                                 : int  18 15 34 52 46 42 51 26 45 44 ...
##  $ Number.of.sexual.partners           : int  4 1 1 5 3 3 3 1 1 3 ...
##  $ First.sexual.intercourse            : int  15 14 NA 16 21 23 17 26 20 15 ...
##  $ Num.of.pregnancies                  : int  1 1 1 4 4 2 6 3 5 NA ...
##  $ Smokes                              : int  0 0 0 1 0 0 1 0 0 1 ...
##  $ Smokes..years.                      : num  0 0 0 37 0 ...
##  $ Smokes..packs.year.                 : num  0 0 0 37 0 0 3.4 0 0 2.8 ...
##  $ Hormonal.Contraceptives             : int  0 0 0 1 1 0 0 1 0 0 ...
##  $ Hormonal.Contraceptives..years.     : num  0 0 0 3 15 0 0 2 0 0 ...
##  $ IUD                                 : int  0 0 0 0 0 0 1 1 0 NA ...
##  $ IUD..years.                         : num  0 0 0 0 0 0 7 7 0 NA ...
##  $ STDs                                : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ STDs..number.                       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ STDs.condylomatosis                 : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ STDs.cervical.condylomatosis        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ STDs.vaginal.condylomatosis         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ STDs.vulvo.perineal.condylomatosis: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ STDs.syphilis                       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ STDs.pelvic.inflammatory.disease    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ STDs.genital.herpes                 : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ STDs.molluscum.contagiosum          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ STDs.AIDS                           : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ STDs.HIV                            : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ STDs.Hepatitis.B                    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ STDs.HPV                            : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ STDs..Number.of.diagnosis           : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ STDs..Time.since.first.diagnosis    : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ STDs..Time.since.last.diagnosis     : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ Dx.Cancer                           : int  0 0 0 1 0 0 0 0 1 0 ...
##  $ Dx.CIN                              : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Dx.HPV                              : int  0 0 0 1 0 0 0 0 1 0 ...
##  $ Dx                                  : int  0 0 0 0 0 0 0 0 1 0 ...
##  $ Hinselmann                          : int  0 0 0 0 0 0 1 0 0 0 ...
##  $ Schiller                            : int  0 0 0 0 0 0 1 0 0 0 ...
##  $ Citology                            : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Biopsy                              : int  0 0 0 0 0 0 1 0 0 0 ...
```

```
dim(cervicalCA)
```

```
## [1] 858  36
```

```
class(cervicalCA)
```

```
## [1] "data.frame"
```

```
glimpse(cervicalCA)
```

```
## Rows: 858
## Columns: 36
## $ Age                                <int> 18, 15, 34, 52, 46, 42, 51, 26, 45,~
## $ Number.of.sexual.partners          <int> 4, 1, 1, 5, 3, 3, 3, 1, 1, 3, 3, 1,~
## $ First.sexual.intercourse           <int> 15, 14, NA, 16, 21, 23, 17, 26, 20,~
## $ Num.of.pregnancies                 <int> 1, 1, 1, 4, 4, 2, 6, 3, 5, NA, 4, 3~
## $ Smokes                             <int> 0, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 0,~
## $ Smokes..years.                     <dbl> 0.000000, 0.000000, 0.000000, 37.00~
## $ Smokes..packs.year.                <dbl> 0.0, 0.0, 0.0, 37.0, 0.0, 0.0, 3.4,~
## $ Hormonal.Contraceptives            <int> 0, 0, 0, 1, 1, 0, 0, 1, 0, 0, 1, 1,~
## $ Hormonal.Contraceptives..years.    <dbl> 0.00, 0.00, 0.00, 3.00, 15.00, 0.00~
## $ IUD                                <int> 0, 0, 0, 0, 0, 0, 1, 1, 0, NA, 0, 0~
## $ IUD..years.                        <dbl> 0, 0, 0, 0, 0, 0, 7, 7, 0, NA, 0, 0~
## $ STDs                               <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ STDs..number.                      <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ STDs.condylomatosis                <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ STDs.cervical.condylomatosis       <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ STDs.vaginal.condylomatosis        <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ STDs.vulvo.perineal.condylomatosis <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ STDs.syphilis                      <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ STDs.pelvic.inflammatory.disease   <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ STDs.genital.herpes                <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ STDs.molluscum.contagiosum         <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ STDs.AIDS                          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ STDs.HIV                           <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ STDs.Hepatitis.B                   <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ STDs.HPV                           <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ STDs..Number.of.diagnosis          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ STDs..Time.since.first.diagnosis   <int> NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ STDs..Time.since.last.diagnosis    <int> NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ Dx.Cancer                          <int> 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0,~
## $ Dx.CIN                             <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ Dx.HPV                             <int> 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0,~
## $ Dx                                 <int> 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0,~
## $ Hinselmann                         <int> 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0,~
## $ Schiller                           <int> 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0,~
## $ Citology                           <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ Biopsy                             <int> 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0,~
```

```
head(cervicalCA)
```

```
##   Age Number.of.sexual.partners First.sexual.intercourse Num.of.pregnancies
## 1  18                         4                       15                 1
## 2  15                         1                       14                 1
## 3  34                         1                       NA                 1
## 4  52                         5                       16                 4
## 5  46                         3                       21                 4
## 6  42                         3                       23                 2
##   Smokes Smokes..years. Smokes..packs.year. Hormonal.Contraceptives
## 1      0              0                   0                       0
## 2      0              0                   0                       0
## 3      0              0                   0                       0
## 4      1             37                  37                       1
## 5      0              0                   0                       1
## 6      0              0                   0                       0
##   Hormonal.Contraceptives..years. IUD IUD..years. STDs STDs..number.
## 1                               0   0           0    0             0
## 2                               0   0           0    0             0
## 3                               0   0           0    0             0
## 4                               3   0           0    0             0
## 5                              15   0           0    0             0
## 6                               0   0           0    0             0
##   STDs.condylomatosis STDs.cervical.condylomatosis STDs.vaginal.condylomatosis
## 1                   0                            0                          0
## 2                   0                            0                          0
## 3                   0                            0                          0
## 4                   0                            0                          0
## 5                   0                            0                          0
## 6                   0                            0                          0
##   STDs.vulvo.perineal.condylomatosis STDs.syphilis
## 1                                  0             0
## 2                                  0             0
## 3                                  0             0
## 4                                  0             0
## 5                                  0             0
## 6                                  0             0
##   STDs.pelvic.inflammatory.disease STDs.genital.herpes
## 1                                0                   0
## 2                                0                   0
## 3                                0                   0
## 4                                0                   0
## 5                                0                   0
## 6                                0                   0
##   STDs.molluscum.contagiosum STDs.AIDS STDs.HIV STDs.Hepatitis.B STDs.HPV
## 1                          0         0        0                0        0
## 2                          0         0        0                0        0
## 3                          0         0        0                0        0
## 4                          0         0        0                0        0
## 5                          0         0        0                0        0
## 6                          0         0        0                0        0
##   STDs..Number.of.diagnosis STDs..Time.since.first.diagnosis
## 1                         0                              NA
## 2                         0                              NA
## 3                         0                              NA
## 4                         0                              NA
## 5                         0                              NA
## 6                         0                              NA
##   STDs..Time.since.last.diagnosis Dx.Cancer Dx.CIN Dx.HPV Dx Hinselmann
```

```
## 1                            NA         0      0      0  0           0
## 2                            NA         0      0      0  0           0
## 3                            NA         0      0      0  0           0
## 4                            NA         1      0      1  0           0
## 5                            NA         0      0      0  0           0
## 6                            NA         0      0      0  0           0
##   Schiller Citology Biopsy
## 1        0        0      0
## 2        0        0      0
## 3        0        0      0
## 4        0        0      0
## 5        0        0      0
## 6        0        0      0
```

As you can see, these commands provide useful information about the quality of the data as well as what the data is.

Replacing Column Names

Based on the “names” command from earlier, the column names of the data in this dataset is very ambiguous. By using a “dplyr” command “rename” we can rename the columns so that it is easier to read.

```
#
# Cleaning up the column names
names(cervicalCA) <- c("age", "sexual_partners","first_intercourse", "pregnancies", "smokes", "smoking_t
ime", "smoke_rate", "hormomal_contraceptives", "hormonal_contraceptives_time", "iud", "iud_time", "std",
"std_time", "condylomatosis", "cervical_condylomatosis", "vaginal_condylomatosis", "vulvo_perineal_condy
lomatosis", "syphilis", "pelvic_inflammatory_disease", "genital_herpes", "molluscum_contagiosm", "aids",
"hiv", "hepatitis_b", "hpv", "diagnosis_num", "first_diagnosis_time", "last_diagnosis_time", "dx_cancer"
, "dx_cin", "dx_hpv", "dx", "hinselmann", "schiller", "cytology", "biopsy")
```

Though this looks complicated, all this is doing is converting the old ambiguous column names to something more understandable.

```
#
# renamed columns
names(cervicalCA)
```

```
##  [1] "age"                           "sexual_partners"
##  [3] "first_intercourse"             "pregnancies"
##  [5] "smokes"                        "smoking_time"
##  [7] "smoke_rate"                    "hormomal_contraceptives"
##  [9] "hormonal_contraceptives_time"  "iud"
## [11] "iud_time"                      "std"
## [13] "std_time"                      "condylomatosis"
## [15] "cervical_condylomatosis"       "vaginal_condylomatosis"
## [17] "vulvo_perineal_condylomatosis" "syphilis"
## [19] "pelvic_inflammatory_disease"   "genital_herpes"
## [21] "molluscum_contagiosm"          "aids"
## [23] "hiv"                           "hepatitis_b"
## [25] "hpv"                           "diagnosis_num"
## [27] "first_diagnosis_time"          "last_diagnosis_time"
## [29] "dx_cancer"                     "dx_cin"
## [31] "dx_hpv"                        "dx"
## [33] "hinselmann"                    "schiller"
## [35] "cytology"                      "biopsy"
```

Checking for Missing Data

Most columns have missing data, either as “NA” or data that just doesn’t make sense. I can check for missing data using the commands below which print the rows with missing data. After knowing that I have missing data, I can run the command below which tells me which columns have missing data.

```
#
# Check for missing data
clean <- ifelse(complete.cases(cervicalCA) == TRUE, 1, 0)
paste("There are", dim(cervicalCA)[1]-sum(clean), "rows with missing data.")
```

```
## [1] "There are 799 rows with missing data."
```

```
#
# Find which columns have missing data
missingDataCol <-colnames(cervicalCA)[apply(cervicalCA, 2, anyNA)]
paste0("The following columns have missing data: ", missingDataCol)
```

```
##  [1] "The following columns have missing data: sexual_partners"
##  [2] "The following columns have missing data: first_intercourse"
##  [3] "The following columns have missing data: pregnancies"
##  [4] "The following columns have missing data: smokes"
##  [5] "The following columns have missing data: smoking_time"
##  [6] "The following columns have missing data: smoke_rate"
##  [7] "The following columns have missing data: hormomal_contraceptives"
##  [8] "The following columns have missing data: hormonal_contraceptives_time"
##  [9] "The following columns have missing data: iud"
## [10] "The following columns have missing data: iud_time"
## [11] "The following columns have missing data: std"
## [12] "The following columns have missing data: std_time"
## [13] "The following columns have missing data: condylomatosis"
## [14] "The following columns have missing data: cervical_condylomatosis"
## [15] "The following columns have missing data: vaginal_condylomatosis"
## [16] "The following columns have missing data: vulvo_perineal_condylomatosis"
## [17] "The following columns have missing data: syphilis"
## [18] "The following columns have missing data: pelvic_inflammatory_disease"
## [19] "The following columns have missing data: genital_herpes"
## [20] "The following columns have missing data: molluscum_contagiosm"
## [21] "The following columns have missing data: aids"
## [22] "The following columns have missing data: hiv"
## [23] "The following columns have missing data: hepatitis_b"
## [24] "The following columns have missing data: hpv"
## [25] "The following columns have missing data: first_diagnosis_time"
## [26] "The following columns have missing data: last_diagnosis_time"
```

Replacing Missing Data

Most of this data is Boolean, so the logical replacement for missing data would be “0”. However, for the numerical data, I can replace the missing values with the mean of the existing values.

```
#
# Replacing the missing data with the mean for
# appropriate columns
# Replacing sexual partners data
cervicalCA$sexual_partners <- ifelse(is.na(cervicalCA$sexual_partners), round(mean(cervicalCA$sexual_par
tners, na.rm=TRUE),0), cervicalCA$sexual_partners)
#
# Replacing first_intercourse data
cervicalCA$first_intercourse <- ifelse(is.na(cervicalCA$first_intercourse), round(mean(cervicalCA$first_
intercourse, na.rm=TRUE),0), cervicalCA$first_intercourse)
#
# Replacing pregnancies data
cervicalCA$pregnancies <- ifelse(is.na(cervicalCA$pregnancies), round(mean(cervicalCA$pregnancies, na.rm
=TRUE),0), cervicalCA$pregnancies)
```

Now I can replace the Boolean variables

```
#
# Replacing the other missing data with 0
cervicalCA[is.na(cervicalCA)] = 0
```

Other clean up

After looking at the data after the prior changes, I noticed that some of the columns were in integers instead of Boolean. I also fixed the rounding issues that occurred when I replaced the missing numerical data with the mean.

```
#
# Replacing integer data as boolean where appropriate
logicals <- c("smokes", "iud", "std", "condylomatosis", "vaginal_condylomatosis", "vulvo_perineal_condyl
omatosis", "syphilis", "pelvic_inflammatory_disease", "genital_herpes", "molluscum_contagiosm", "aids",
"hiv", "hepatitis_b", "hpv", "dx_cancer", "dx_cin", "dx_hpv", "dx", "hinselmann", "schiller", "cytology"
, "biopsy")
cervicalCA[logicals] <- lapply(cervicalCA[logicals], as.logical)
#
# Round values in columns with many digits
cervicalCA$smoking_time <- round(cervicalCA$smoking_time, 3)
cervicalCA$smoke_rate <- round(cervicalCA$smoke_rate, 3)
cervicalCA$hormonal_contraceptives_time <- round(cervicalCA$hormonal_contraceptives_time,3)
```

Now I can check my data set before I continue.

```
summary(cervicalCA)
```

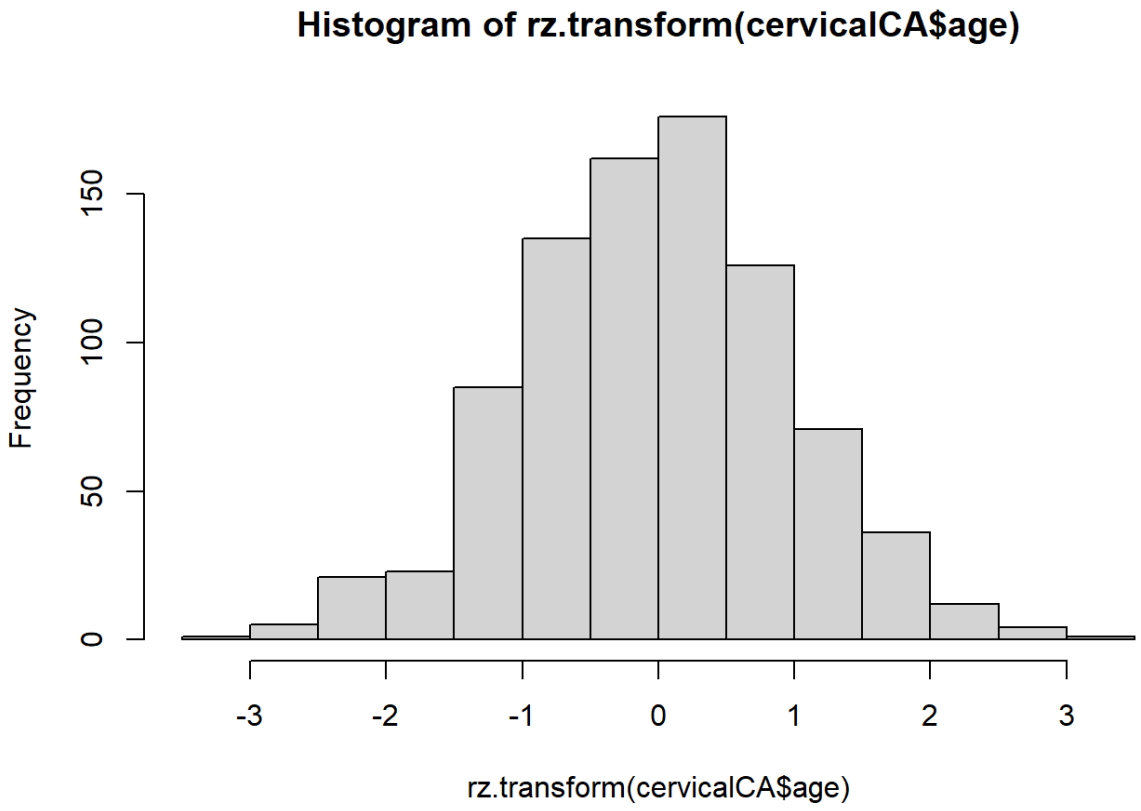
```
##       age        sexual_partners  first_intercourse  pregnancies
##  Min.   :13.00   Min.   : 1.000   Min.   :10        Min.   : 0.000
##  1st Qu.:20.00   1st Qu.: 2.000   1st Qu.:15        1st Qu.: 1.000
##  Median :25.00   Median : 2.000   Median :17        Median : 2.000
##  Mean   :26.82   Mean   : 2.542   Mean   :17        Mean   : 2.258
##  3rd Qu.:32.00   3rd Qu.: 3.000   3rd Qu.:18        3rd Qu.: 3.000
##  Max.   :84.00   Max.   :28.000   Max.   :32        Max.   :11.000
##    smokes         smoking_time      smoke_rate      hormomal_contraceptives
##  Mode :logical   Min.   : 0.000   Min.   : 0.0000   Min.   :0.0000
##  FALSE:735       1st Qu.: 0.000   1st Qu.: 0.0000   1st Qu.:0.0000
##  TRUE :123       Median : 0.000   Median : 0.0000   Median :1.0000
##                  Mean   : 1.201   Mean   : 0.4463   Mean   :0.5606
##                  3rd Qu.: 0.000   3rd Qu.: 0.0000   3rd Qu.:1.0000
##                  Max.   :37.000   Max.   :37.0000   Max.   :1.0000
##  hormonal_contraceptives_time    iud             iud_time          std
##  Min.   : 0.000               Mode :logical   Min.   : 0.0000   Mode :logical
##  1st Qu.: 0.000               FALSE:775       1st Qu.: 0.0000   FALSE:779
##  Median : 0.250               TRUE :83        Median : 0.0000   TRUE :79
##  Mean   : 1.972                               Mean   : 0.4446
##  3rd Qu.: 2.000                               3rd Qu.: 0.0000
##  Max.   :30.000                               Max.   :19.0000
##     std_time     condylomatosis  cervical_condylomatosis vaginal_condylomatosis
##  Min.   :0.000   Mode :logical   Min.   :0               Mode :logical
##  1st Qu.:0.000   FALSE:814       1st Qu.:0               FALSE:854
##  Median :0.000   TRUE :44        Median :0               TRUE :4
##  Mean   :0.155                   Mean   :0
##  3rd Qu.:0.000                   3rd Qu.:0
##  Max.   :4.000                   Max.   :0
##  vulvo_perineal_condylomatosis  syphilis       pelvic_inflammatory_disease
##  Mode :logical                 Mode :logical   Mode :logical
##  FALSE:815                     FALSE:840       FALSE:857
##  TRUE :43                      TRUE :18        TRUE :1
##
##
##
##  genital_herpes  molluscum_contagiosm    aids            hiv
##  Mode :logical   Mode :logical        Mode :logical   Mode :logical
##  FALSE:857       FALSE:857            FALSE:858       FALSE:840
##  TRUE :1         TRUE :1                              TRUE :18
##
##
##
##  hepatitis_b        hpv           diagnosis_num     first_diagnosis_time
##  Mode :logical   Mode :logical   Min.   :0.00000   Min.   : 0.0000
##  FALSE:857       FALSE:856       1st Qu.:0.00000   1st Qu.: 0.0000
##  TRUE :1         TRUE :2         Median :0.00000   Median : 0.0000
##                                  Mean   :0.08741   Mean   : 0.5082
##                                  3rd Qu.:0.00000   3rd Qu.: 0.0000
##                                  Max.   :3.00000   Max.   :22.0000
##  last_diagnosis_time dx_cancer         dx_cin          dx_hpv
##  Min.   : 0.0000     Mode :logical   Mode :logical   Mode :logical
##  1st Qu.: 0.0000     FALSE:840       FALSE:849       FALSE:840
##  Median : 0.0000     TRUE :18        TRUE :9         TRUE :18
##  Mean   : 0.4814
##  3rd Qu.: 0.0000
##  Max.   :22.0000
##      dx          hinselmann       schiller        cytology
```

```
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:834       FALSE:823       FALSE:784       FALSE:814
##  TRUE :24        TRUE :35        TRUE :74        TRUE :44
##
##
##
##    biopsy
##  Mode :logical
##  FALSE:803
##  TRUE :55
##
##
##
```

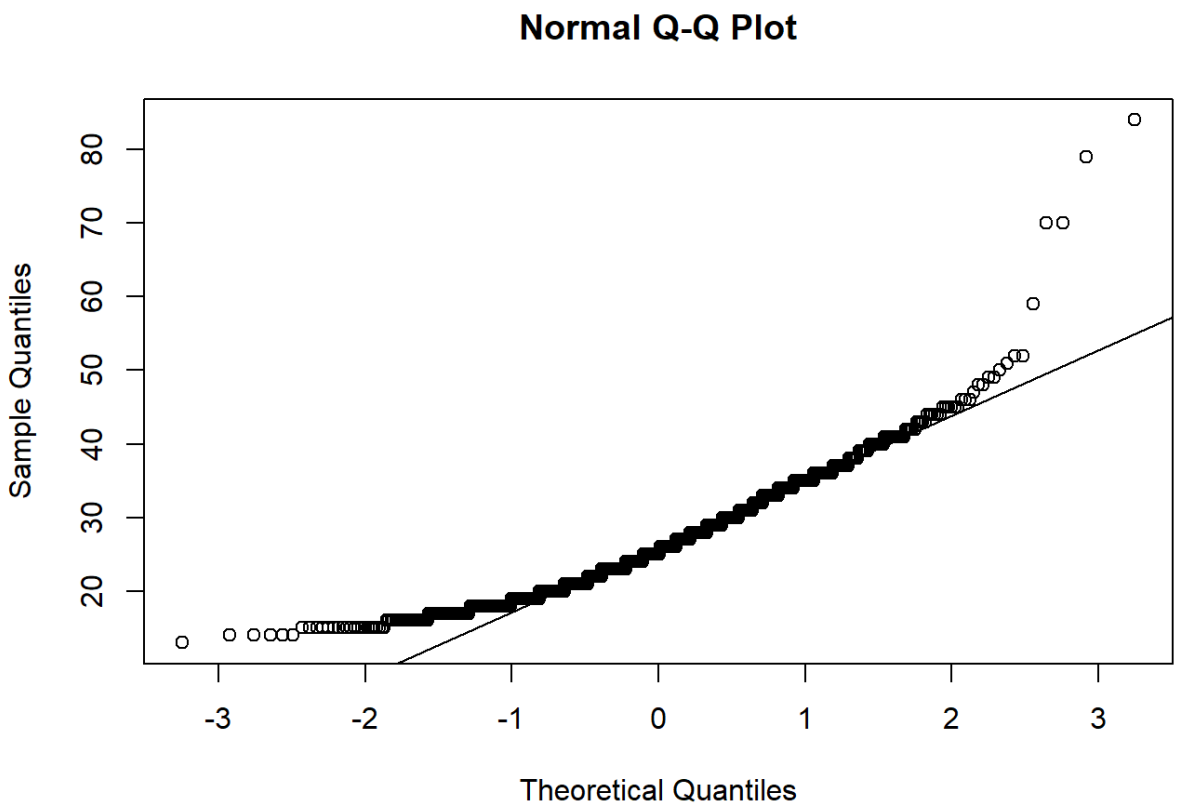
Normalizing data

To get a feel for the data, I need to normalize the data. I used a rank z transform to normalize the data for analysis. I also added a column to the dataset for these values.

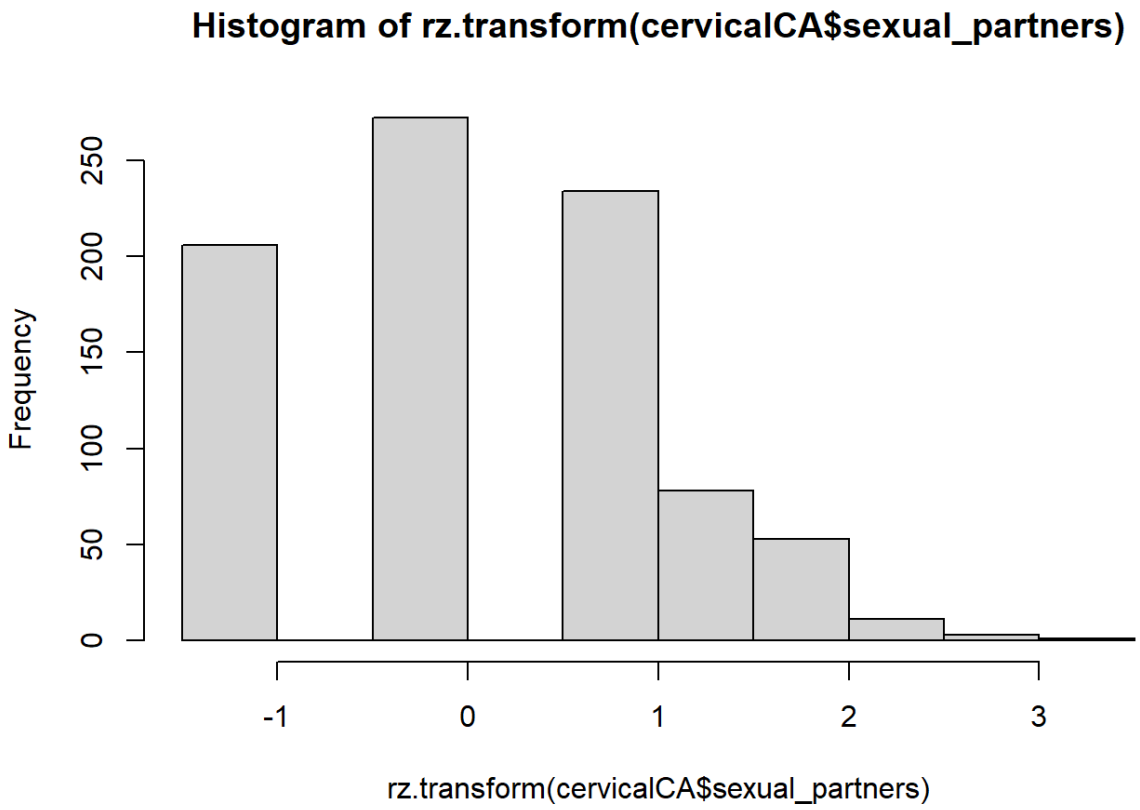
```
#
# Normal distribution of numerical data
# Checking for age
hist(rz.transform(cervicalCA$age))
```



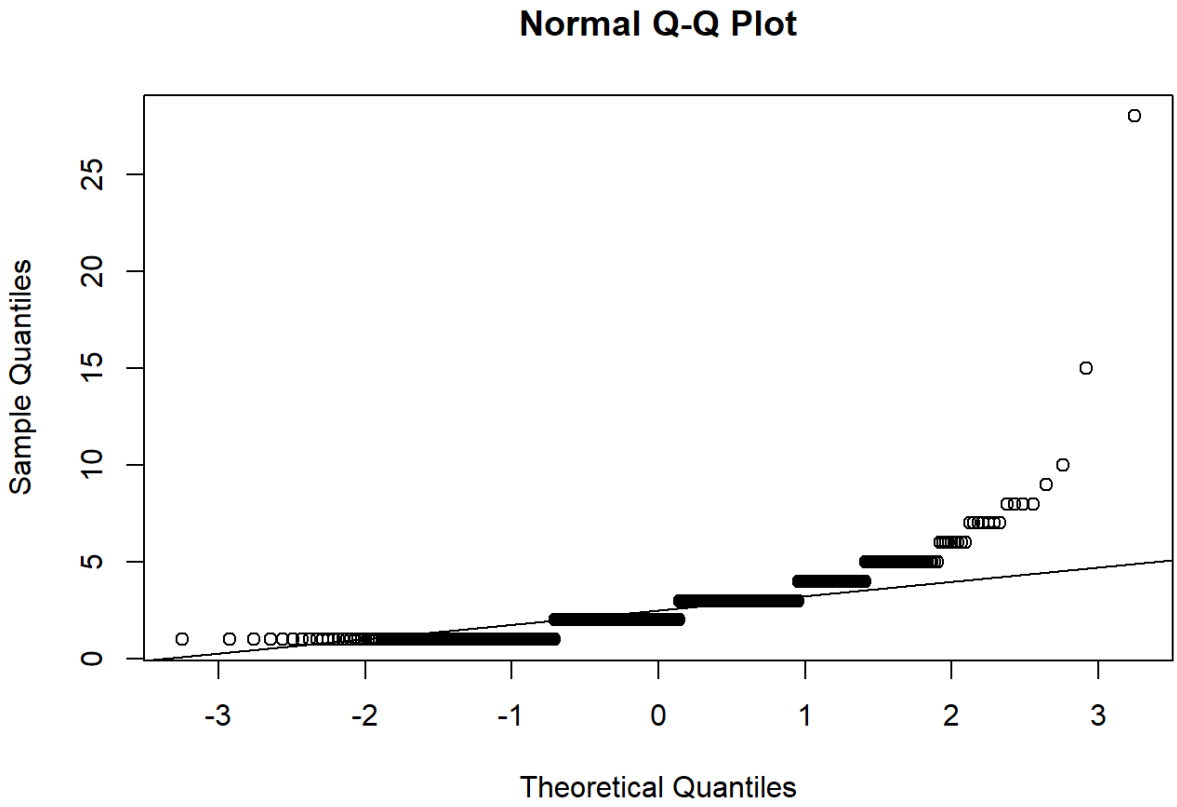

```
qqnorm(cervicalCA$age)
qqline(cervicalCA$age)
```



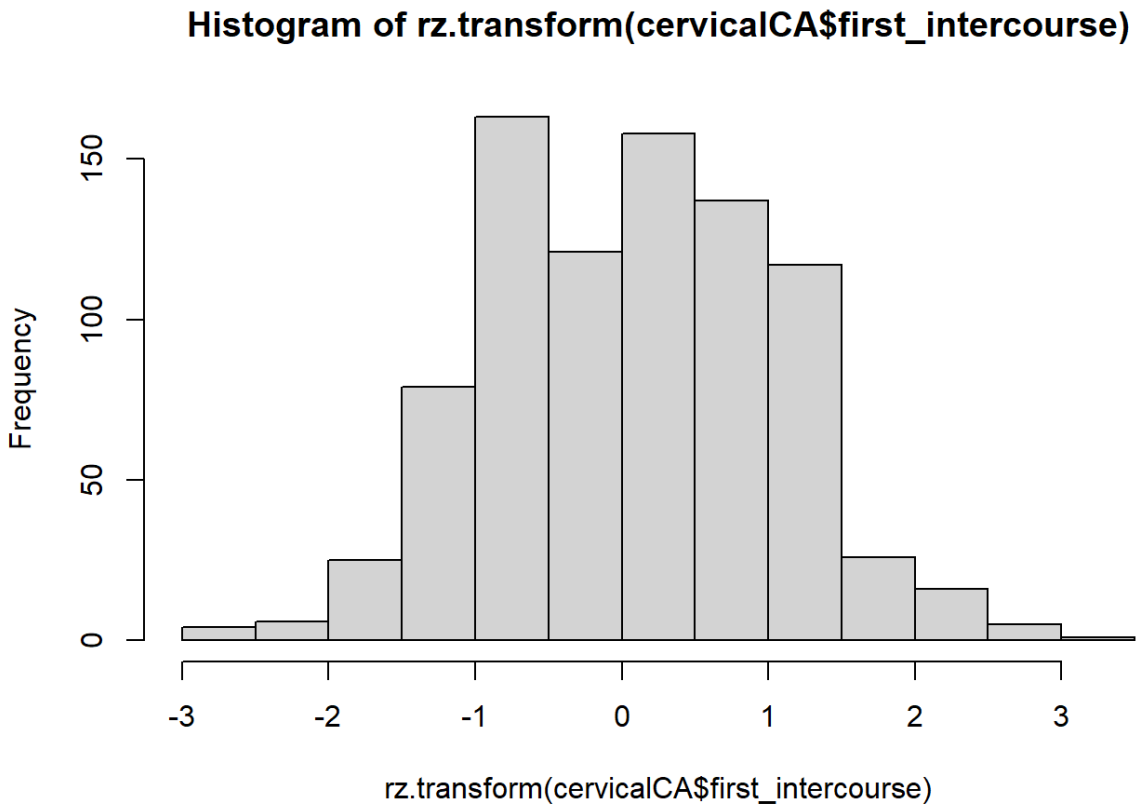

```
cervicalCA$rz_transform_age <- rz.transform(cervicalCA$age)
#
# Checking for sexual_partners
hist(rz.transform(cervicalCA$sexual_partners))
```



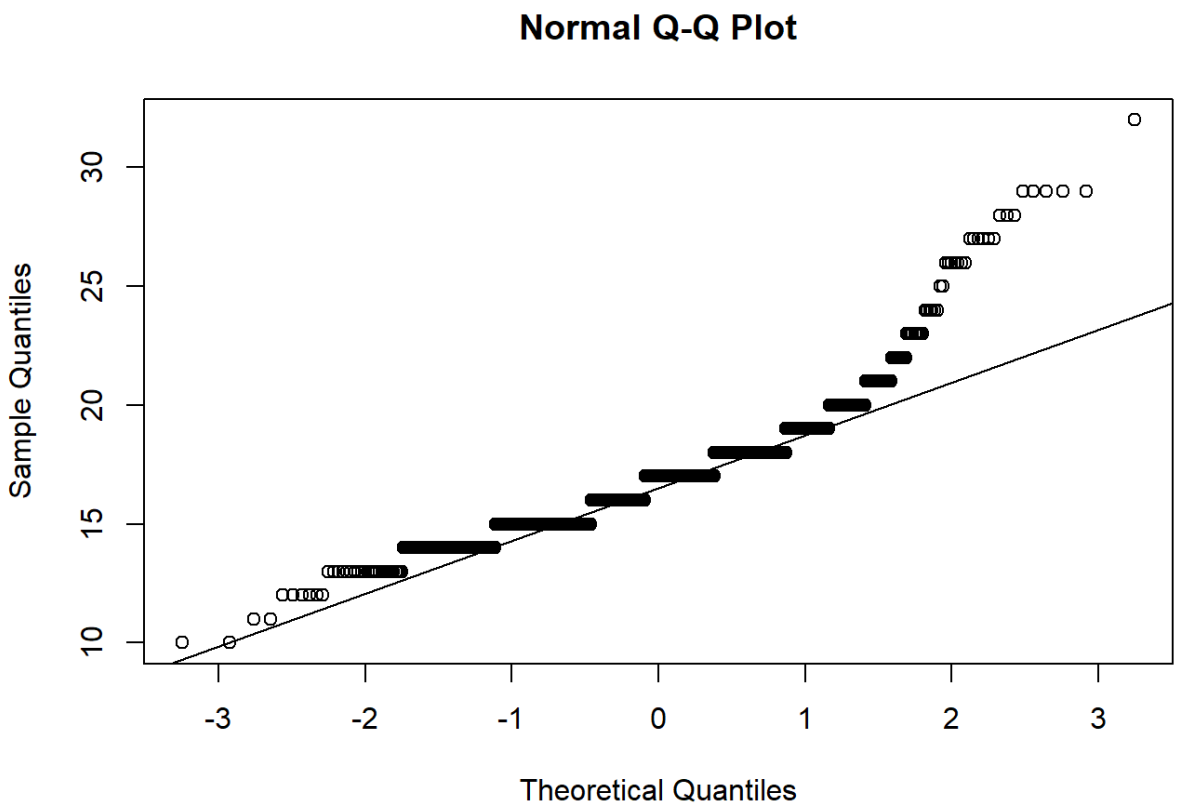

```
qqnorm(cervicalCA$sexual_partners)
qqline(cervicalCA$sexual_partners)
```



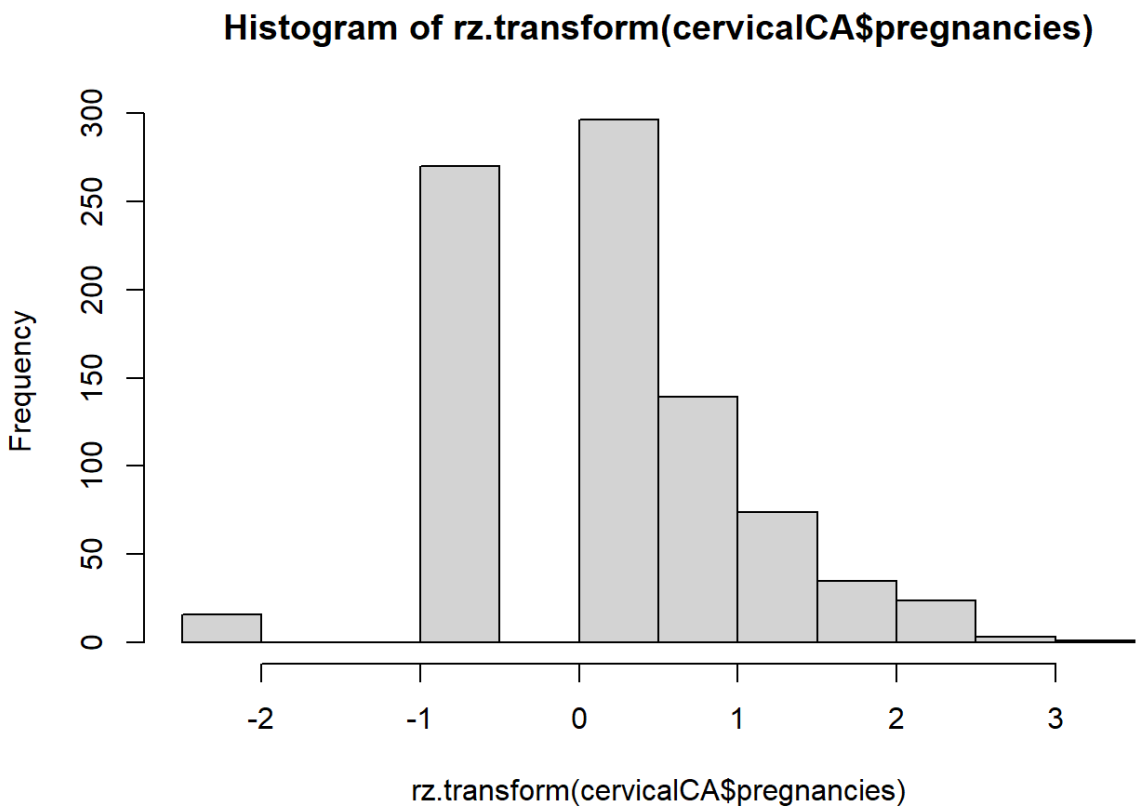
```
cervicalCA$rz_transform_sexual_partners <- rz.transform(cervicalCA$sexual_partners)
#
# Checking for first_intercourse
hist(rz.transform(cervicalCA$first_intercourse))
```



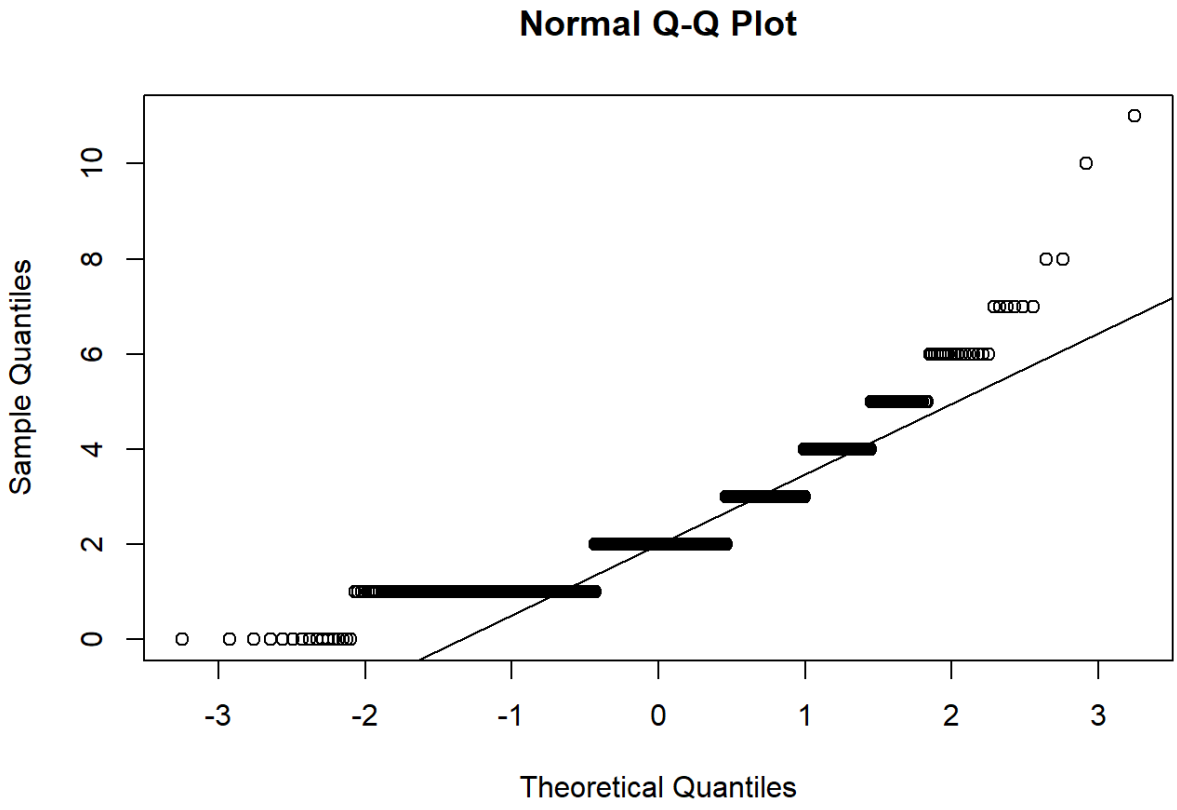

```
qqnorm(cervicalCA$first_intercourse)
qqline(cervicalCA$first_intercourse)
```



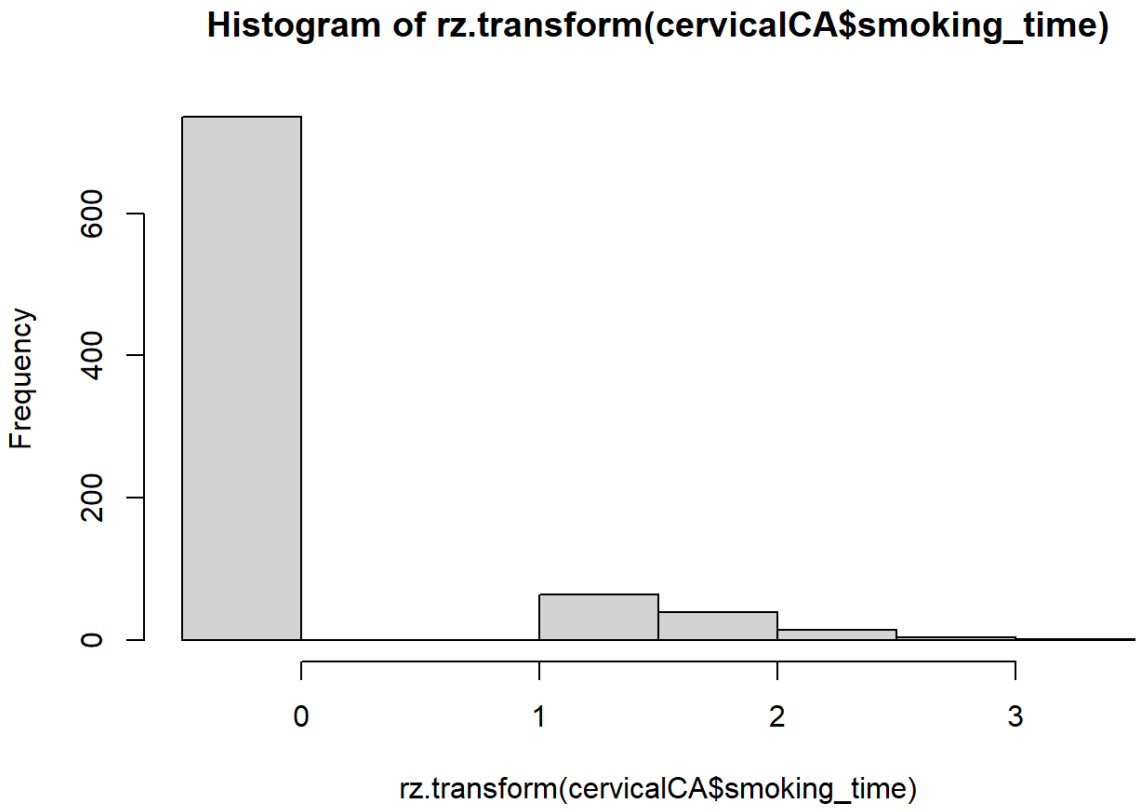

```
cervicalCA$rz_transform_first_intercourse <- rz.transform(cervicalCA$first_intercourse)
#
# Checking for pregnancies
hist(rz.transform(cervicalCA$pregnancies))
```



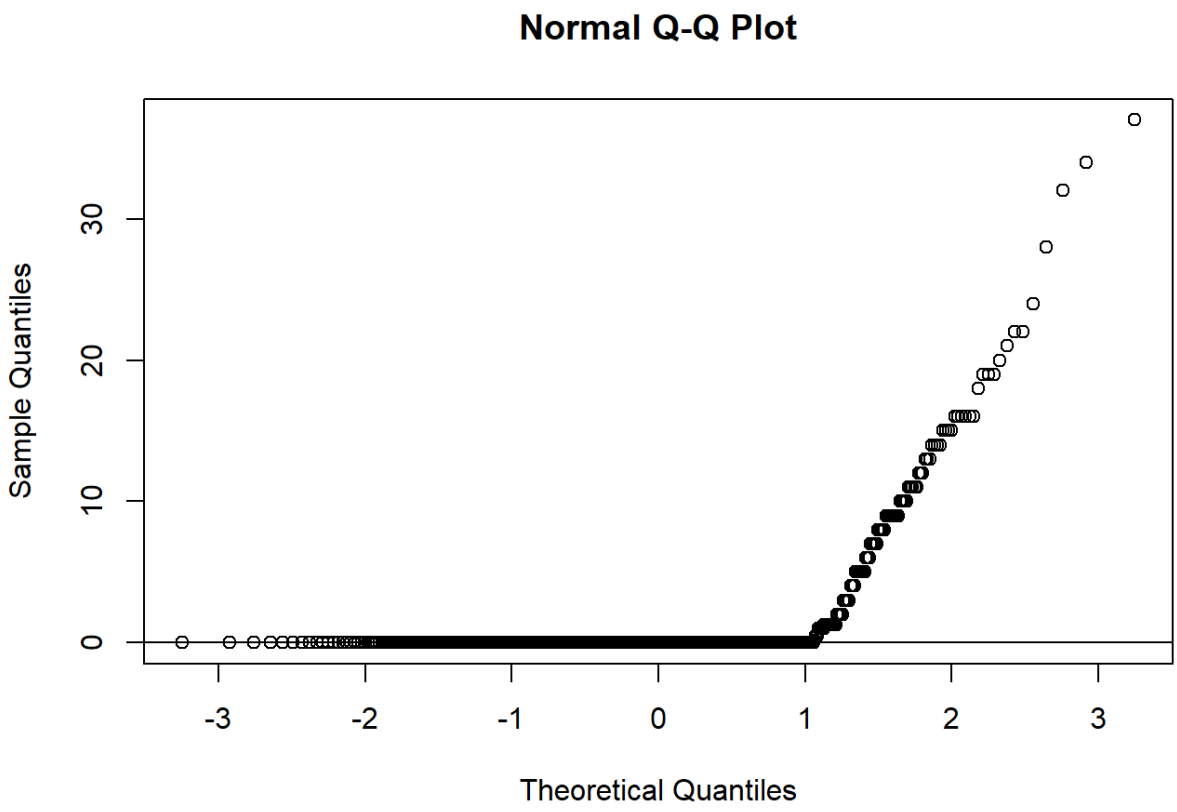

```
qqnorm(cervicalCA$pregnancies)
qqline(cervicalCA$pregnancies)
```



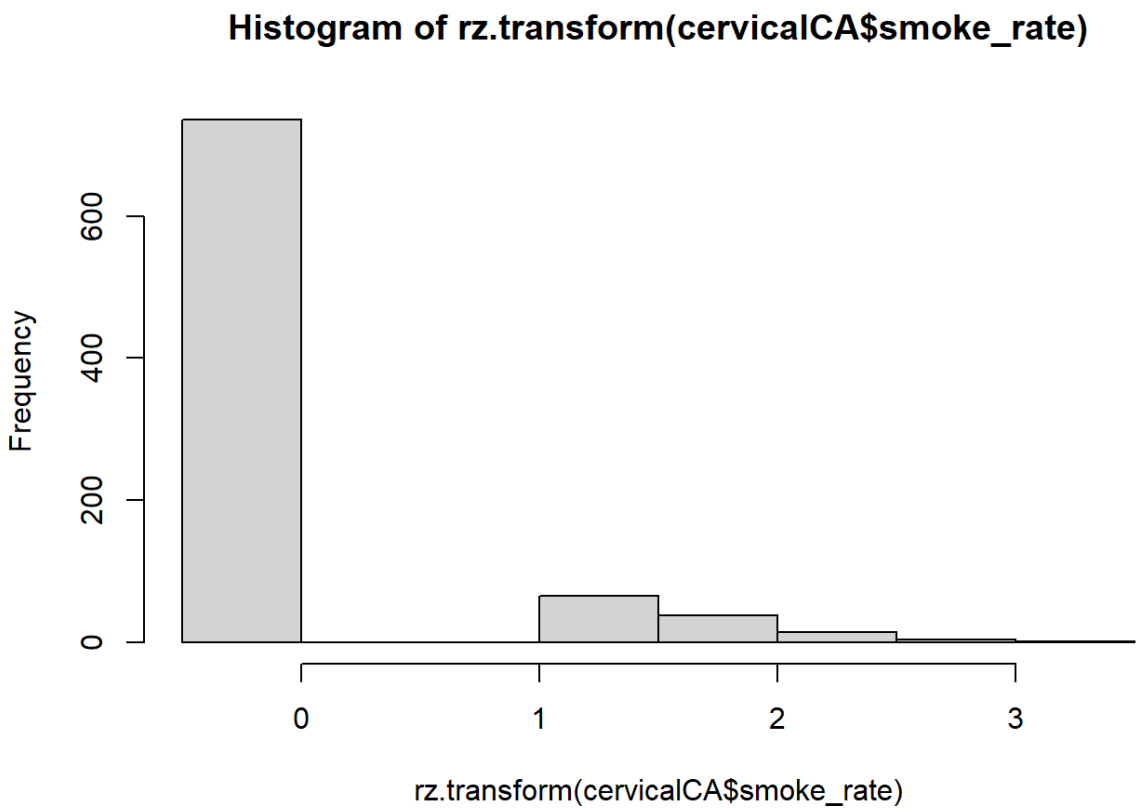
```
cervicalCA$rz_transform_pregnancies <- rz.transform(cervicalCA$pregnancies)
#
# Checking for smoking_time
hist(rz.transform(cervicalCA$smoking_time))
```



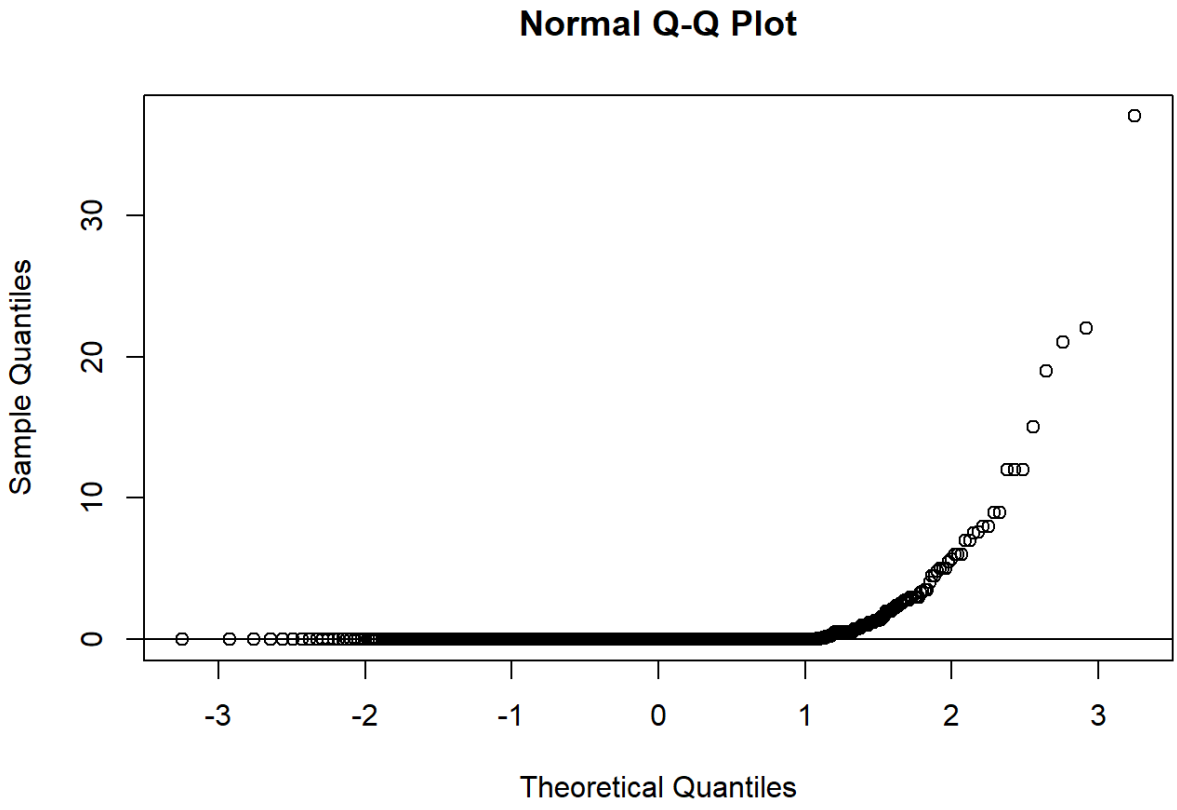

```
qqnorm(cervicalCA$smoking_time)
qqline(cervicalCA$smoking_time)
```



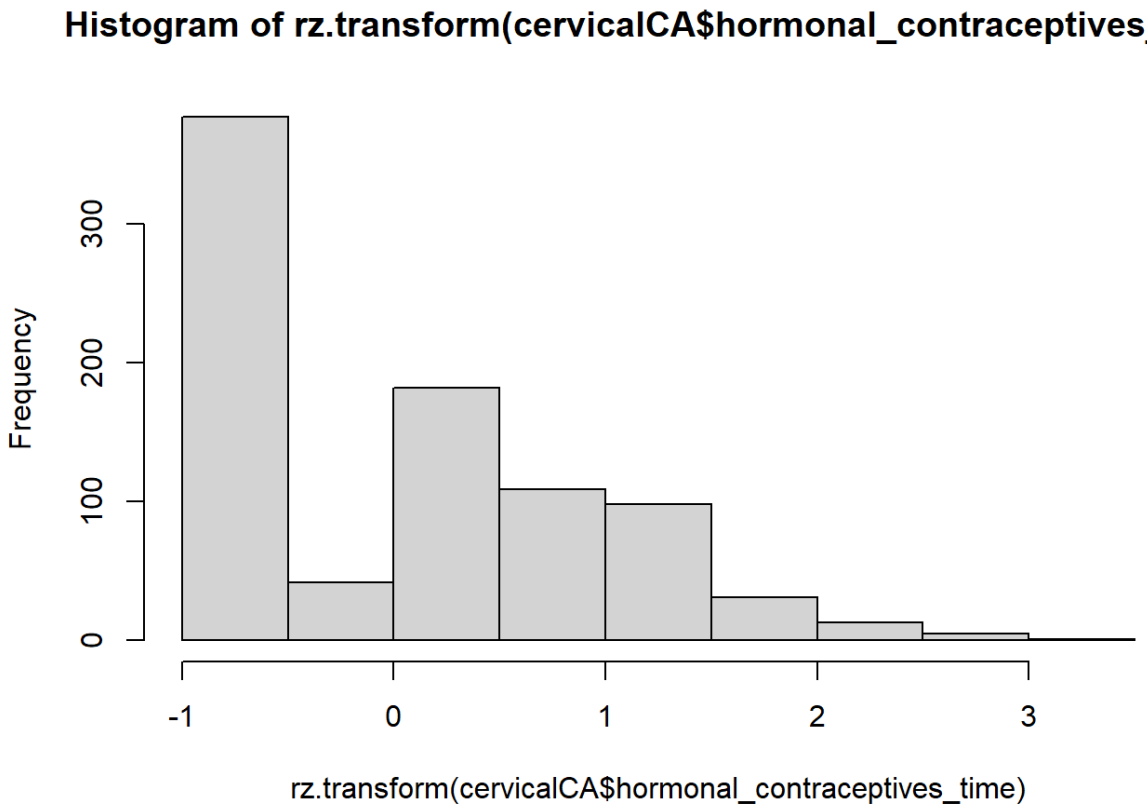

```
cervicalCA$rz_transform_smoking_time <- rz.transform(cervicalCA$smoking_time)
#
# Checking for smoke_rate
hist(rz.transform(cervicalCA$smoke_rate))
```



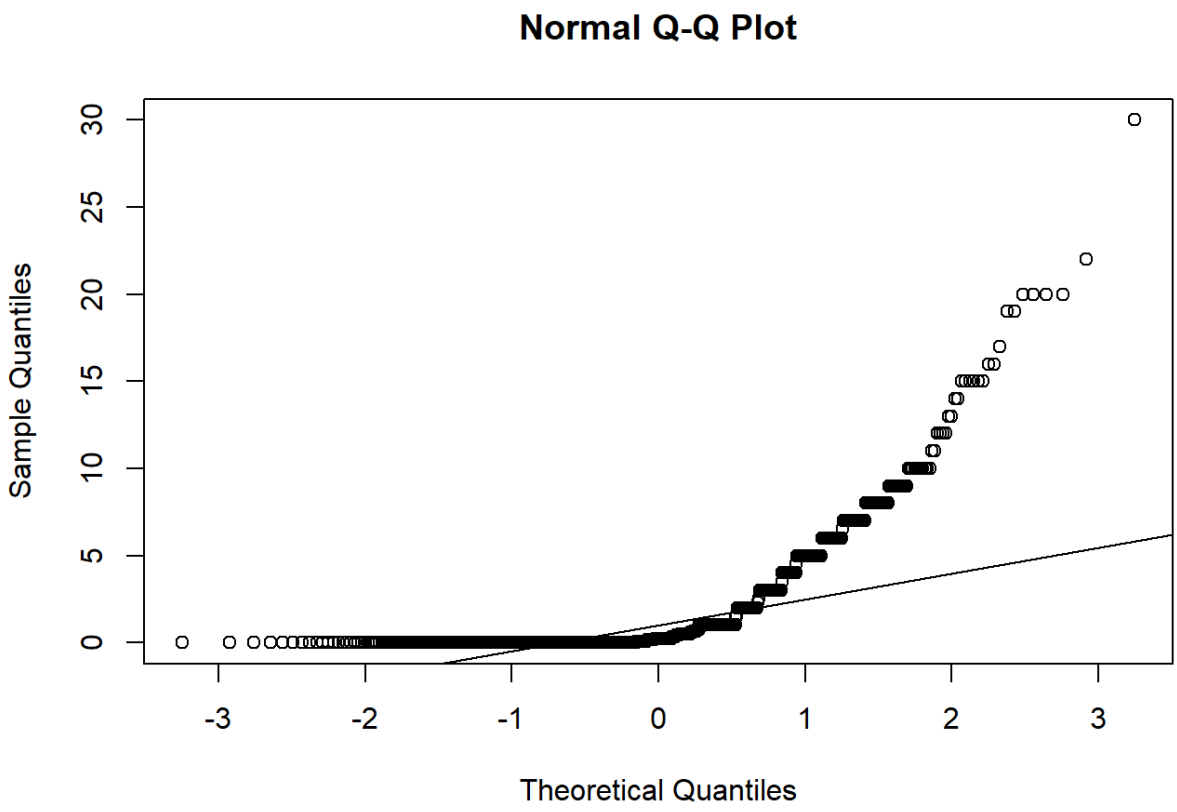

```
qqnorm(cervicalCA$smoke_rate)
qqline(cervicalCA$smoke_rate)
```



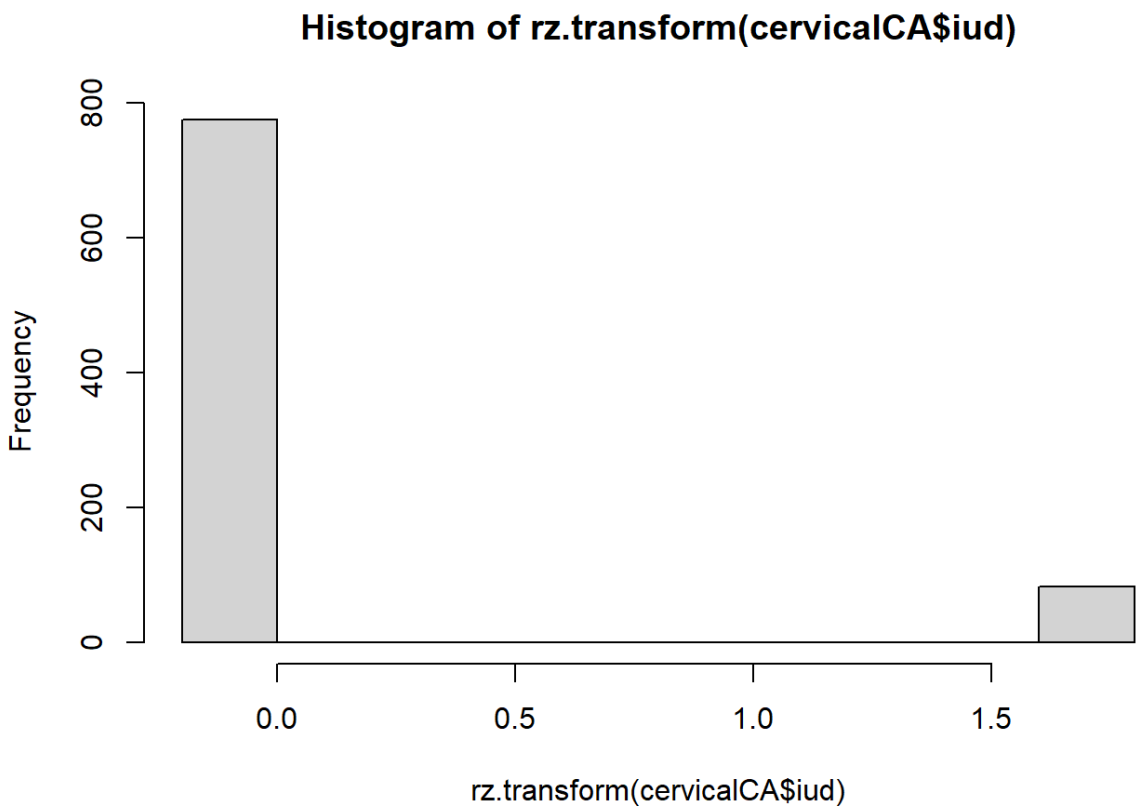
```
cervicalCA$rz_transform_smoke_rate <- rz.transform(cervicalCA$smoke_rate)
#
# Checking for hormonal_contraceptives_time
hist(rz.transform(cervicalCA$hormonal_contraceptives_time))
```



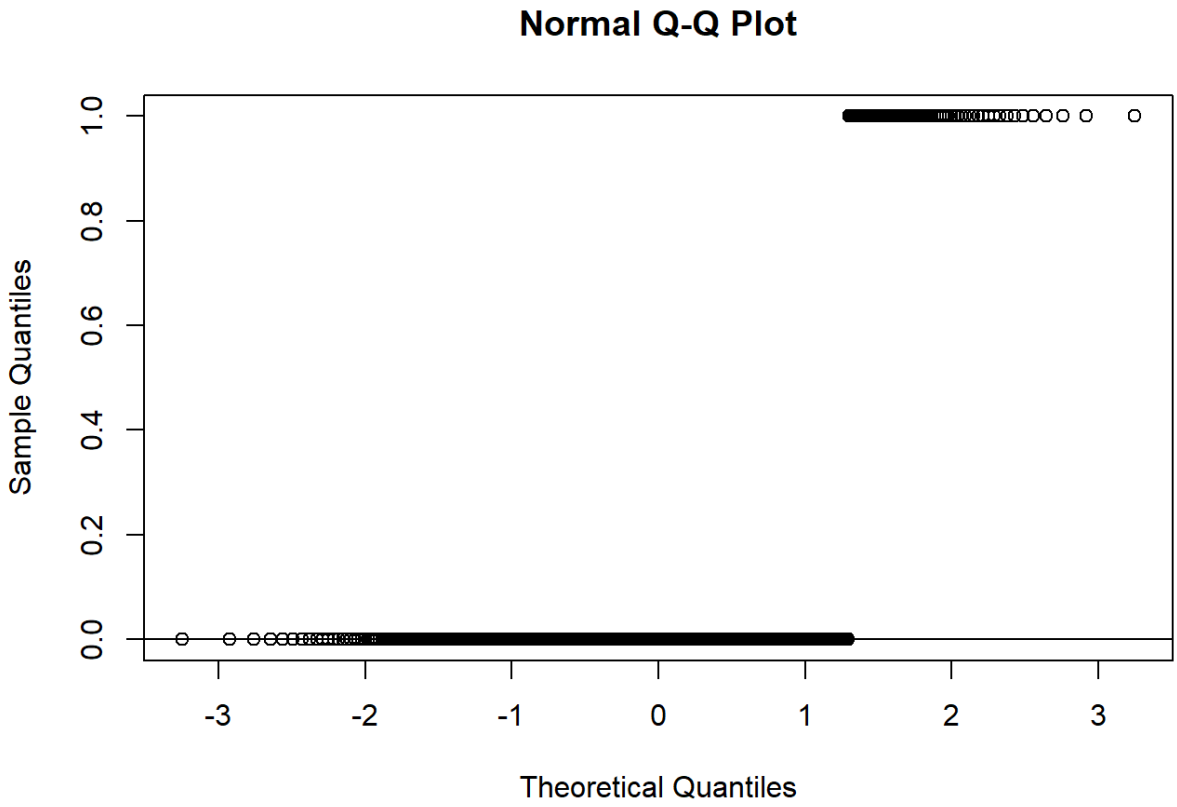

```
qqnorm(cervicalCA$hormonal_contraceptives_time)
qqline(cervicalCA$hormonal_contraceptives_time)
```



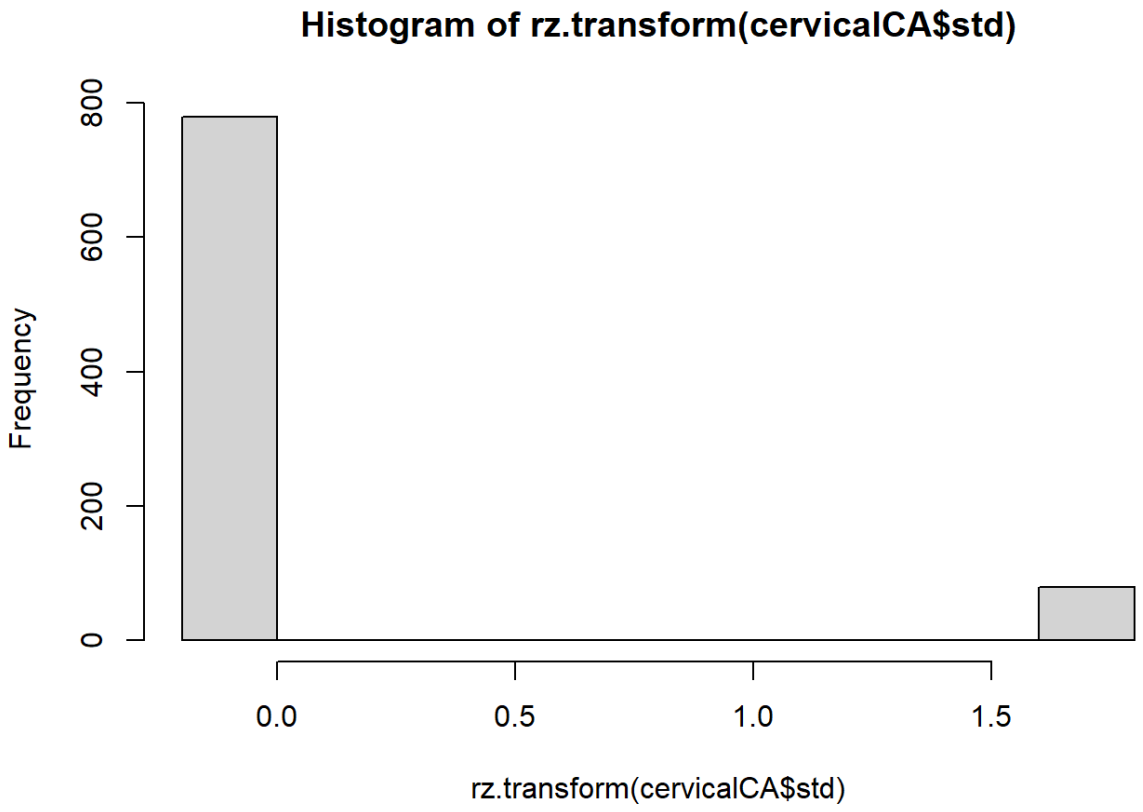

```
cervicalCA$rz_transform_hormonal_contraceptives_time <- rz.transform(cervicalCA$hormonal_contraceptives_
time)
#
# Checking for iud
hist(rz.transform(cervicalCA$iud))
```



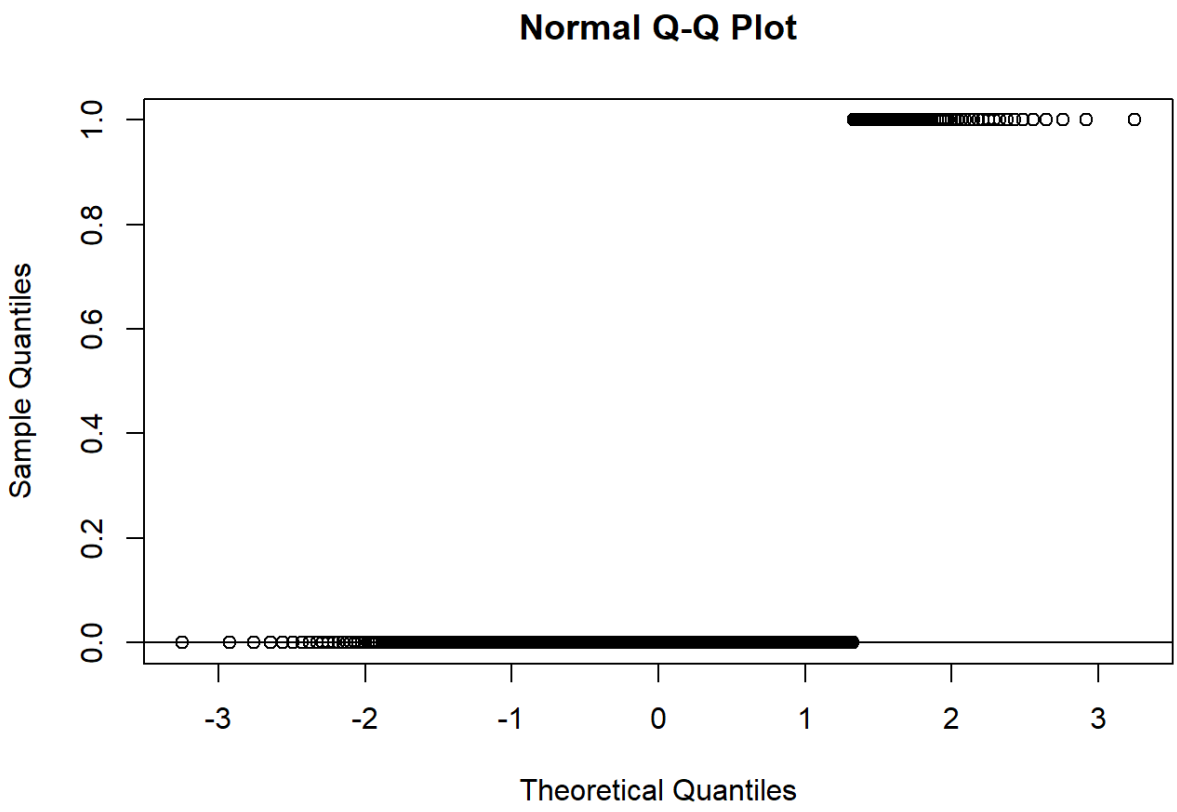

```
qqnorm(cervicalCA$iud)
qqline(cervicalCA$iud)
```



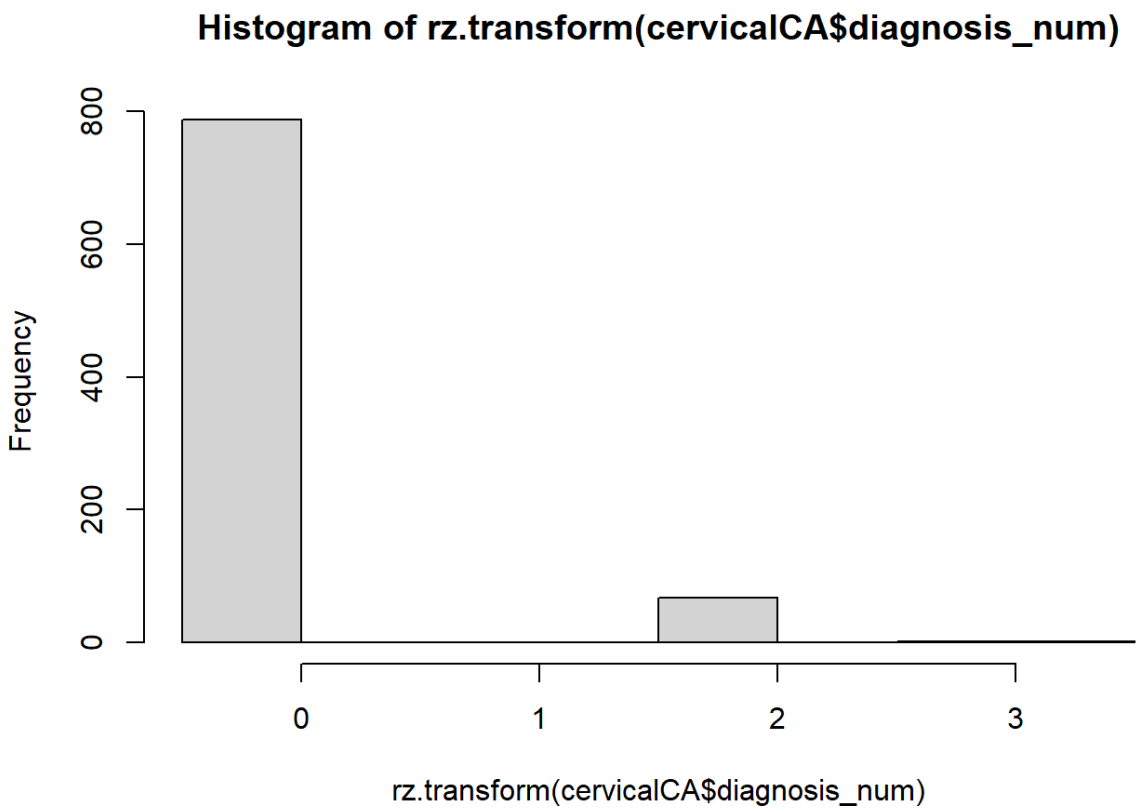
```
cervicalCA$rz_transform_iud <- rz.transform(cervicalCA$iud)
#
# Checking for std
hist(rz.transform(cervicalCA$std))
```



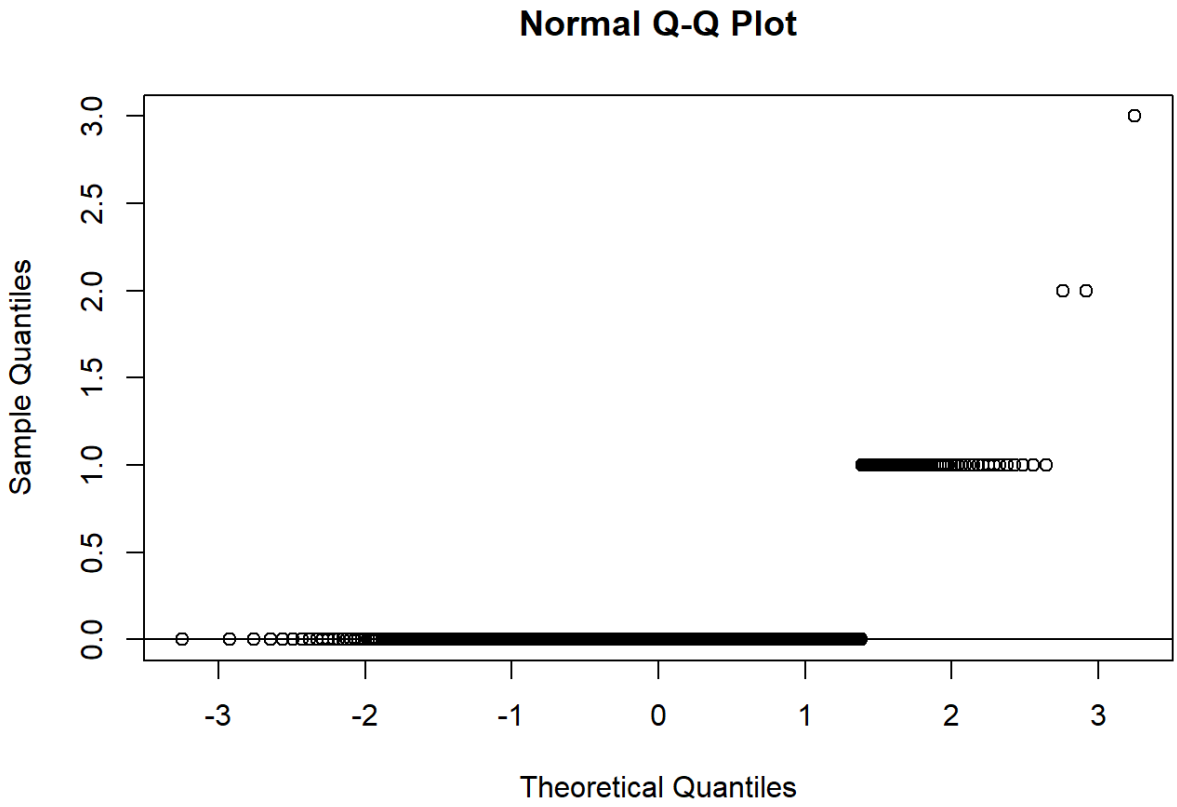

```
qqnorm(cervicalCA$std)
qqline(cervicalCA$std)
```



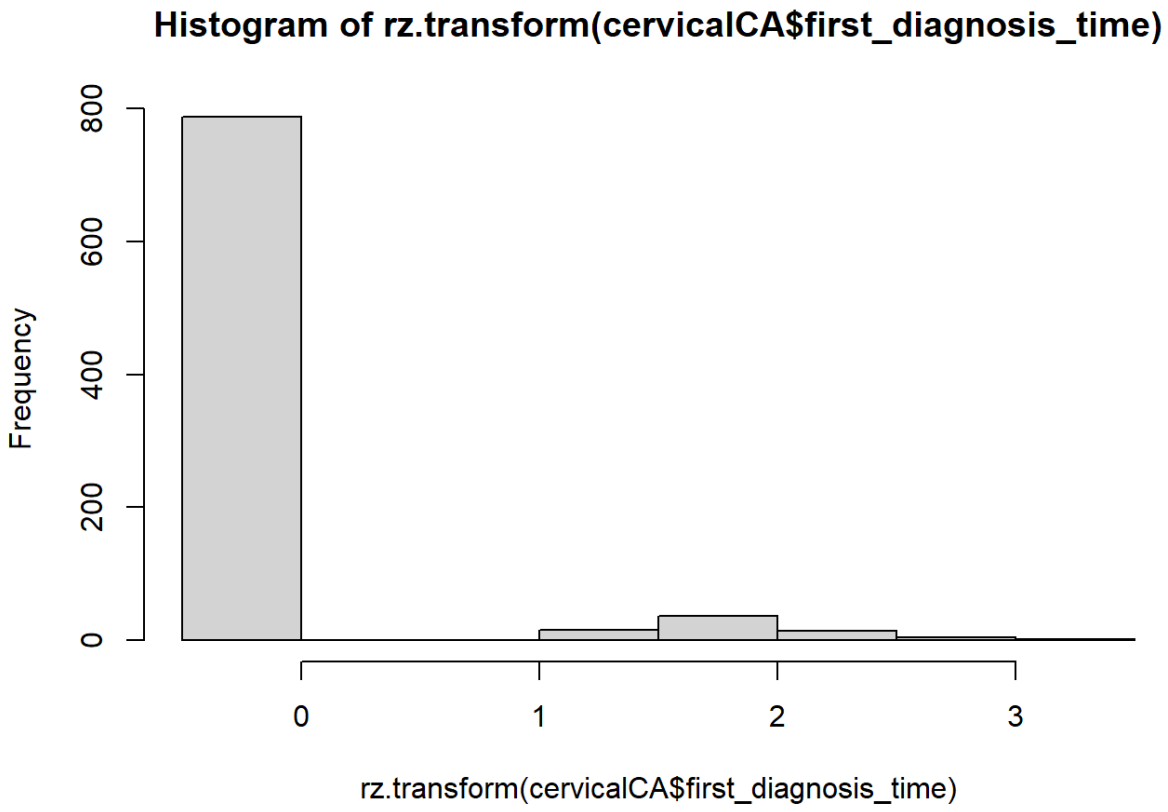

```
cervicalCA$rz_transform_std <- rz.transform(cervicalCA$std)
#
# Checking for diagnosis_num
hist(rz.transform(cervicalCA$diagnosis_num))
```



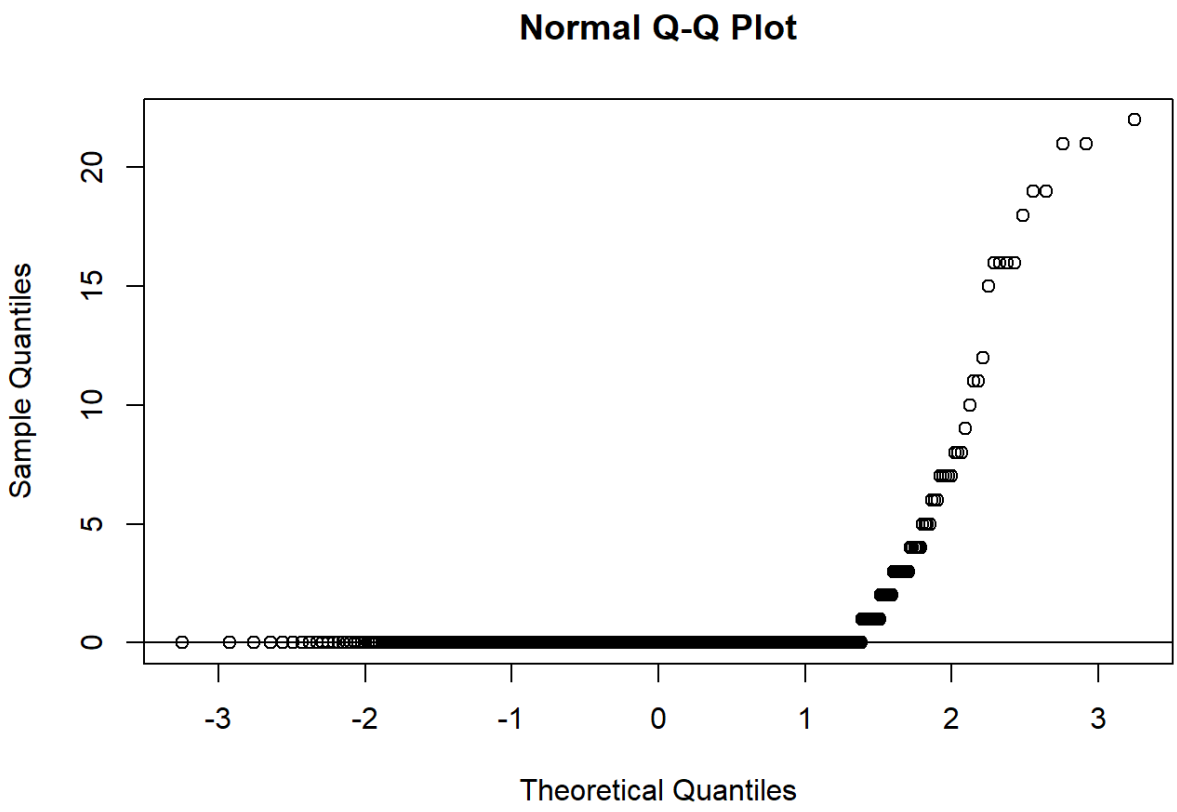

```
qqnorm(cervicalCA$diagnosis_num)
qqline(cervicalCA$diagnosis_num)
```



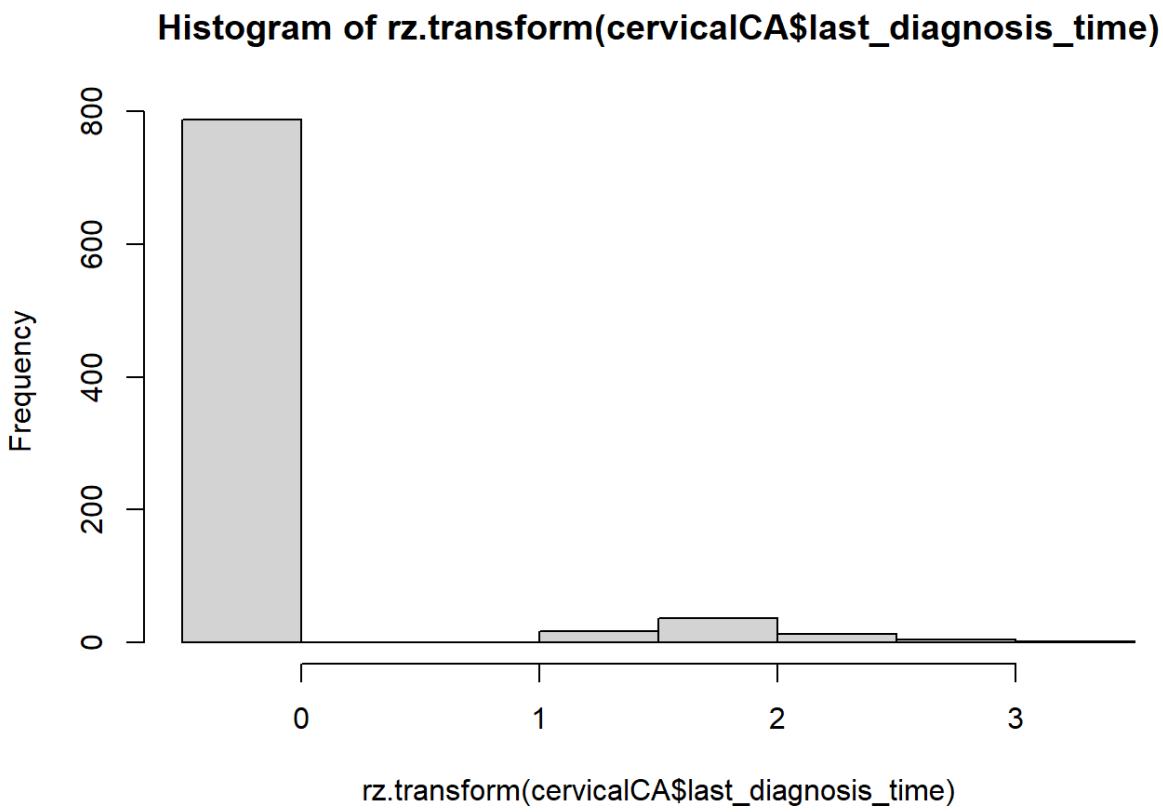
```
cervicalCA$rz_transform_diagnosis_num <- rz.transform(cervicalCA$diagnosis_num)
#
# Checking for first_diagnosis_time
hist(rz.transform(cervicalCA$first_diagnosis_time))
```



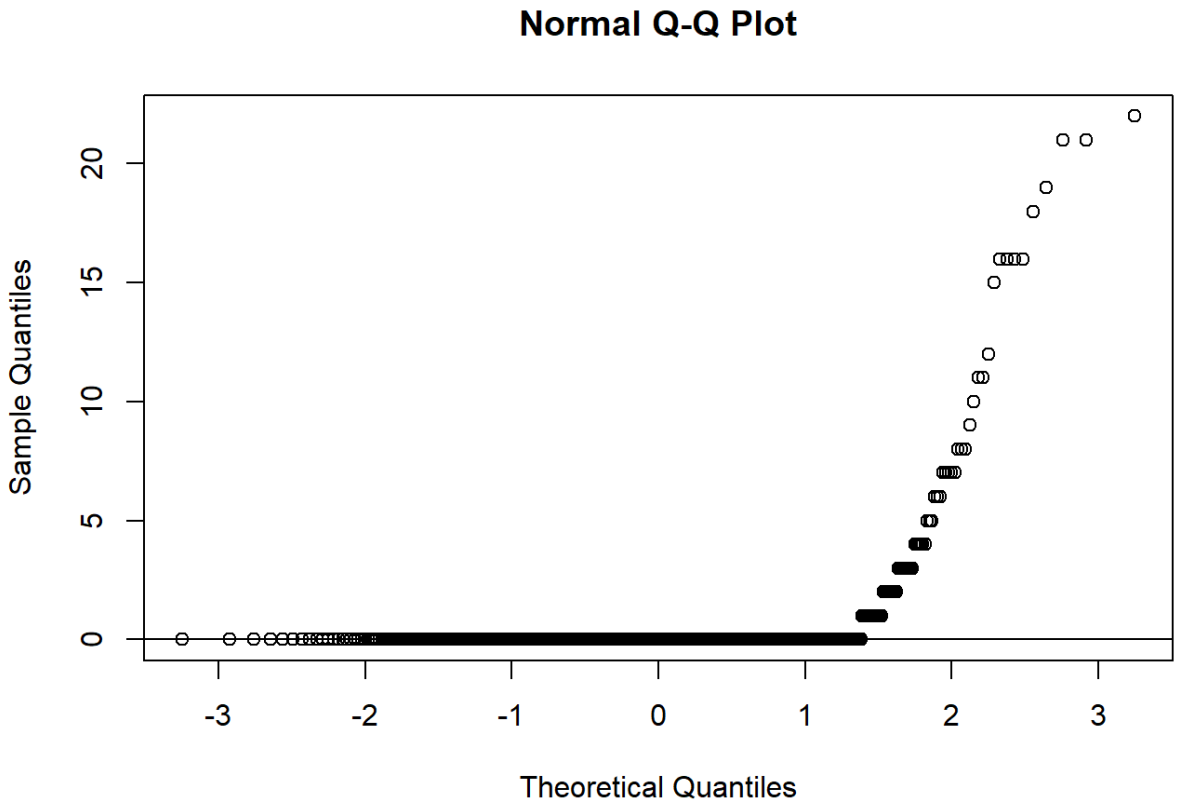

```
qqnorm(cervicalCA$first_diagnosis_time)
qqline(cervicalCA$first_diagnosis_time)
```




```
cervicalCA$rz_transform_first_diagnosis_time <- rz.transform(cervicalCA$first_diagnosis_time)
#
# Checking for last_diagnosis_time
hist(rz.transform(cervicalCA$last_diagnosis_time))
```




```
qqnorm(cervicalCA$last_diagnosis_time)
qqline(cervicalCA$last_diagnosis_time)
```



```
cervicalCA$rz_transform_last_diagnosis_time <- rz.transform(cervicalCA$last_diagnosis_time)
```

Here I export the clean the data to an external file so that I can analyze the data. I have already exported this file so I put the code as a comment.

```
#
# Writing export file
# write.csv(cervicalCA,"/Users/advai/Documents/DSFS/ccdataMod.csv")
```

Loadin the Clean Data

I do the same steps to prepare the work space, load the data, and look at the data.

```
#
# clean up and set up
rm(list=ls())
setwd("/Users/advai/Documents/DSFS")
#
# install and load libraries
library(tidyverse)
library(stringr)
source("myfunctions.R")
#
# load clean Cervical cancer data
cervicalCA <- read.csv(file = "C:\\Users\\advai\\Documents\\DSFS\\ccdataMod.csv",header=T, na.strings=c(
"?"))
#
# looking at the data
names(cervicalCA)
```

```
##  [1] "X"
##  [2] "age"
##  [3] "sexual_partners"
##  [4] "first_intercourse"
##  [5] "pregnancies"
##  [6] "smokes"
##  [7] "smoking_time"
##  [8] "smoke_rate"
##  [9] "hormomal_contraceptives"
## [10] "hormonal_contraceptives_time"
## [11] "iud"
## [12] "iud_time"
## [13] "std"
## [14] "std_time"
## [15] "condylomatosis"
## [16] "cervical_condylomatosis"
## [17] "vaginal_condylomatosis"
## [18] "vulvo_perineal_condylomatosis"
## [19] "syphilis"
## [20] "pelvic_inflammatory_disease"
## [21] "genital_herpes"
## [22] "molluscum_contagiosm"
## [23] "aids"
## [24] "hiv"
## [25] "hepatitis_b"
## [26] "hpv"
## [27] "diagnosis_num"
## [28] "first_diagnosis_time"
## [29] "last_diagnosis_time"
## [30] "dx_cancer"
## [31] "dx_cin"
## [32] "dx_hpv"
## [33] "dx"
## [34] "hinselmann"
## [35] "schiller"
## [36] "cytology"
## [37] "biopsy"
## [38] "rz_transform_age"
## [39] "rz_transform_sexual_partners"
## [40] "rz_transform_first_intercourse"
## [41] "rz_transform_pregnancies"
## [42] "rz_transform_smoking_time"
## [43] "rz_transform_smoke_rate"
## [44] "rz_transform_hormonal_contraceptives_time"
## [45] "rz_transform_iud"
## [46] "rz_transform_std"
## [47] "rz_transform_diagnosis_num"
## [48] "rz_transform_first_diagnosis_time"
## [49] "rz_transform_last_diagnosis_time"
```

```
summary(cervicalCA)
```

```
##        X              age        sexual_partners  first_intercourse
##  Min.   :  1.0   Min.   :13.00   Min.   : 1.000   Min.   :10
##  1st Qu.:215.2   1st Qu.:20.00   1st Qu.: 2.000   1st Qu.:15
##  Median :429.5   Median :25.00   Median : 2.000   Median :17
##  Mean   :429.5   Mean   :26.82   Mean   : 2.542   Mean   :17
##  3rd Qu.:643.8   3rd Qu.:32.00   3rd Qu.: 3.000   3rd Qu.:18
##  Max.   :858.0   Max.   :84.00   Max.   :28.000   Max.   :32
##   pregnancies       smokes         smoking_time      smoke_rate
##  Min.   : 0.000   Mode :logical   Min.   : 0.000   Min.   : 0.0000
##  1st Qu.: 1.000   FALSE:735       1st Qu.: 0.000   1st Qu.: 0.0000
##  Median : 2.000   TRUE :123       Median : 0.000   Median : 0.0000
##  Mean   : 2.258                   Mean   : 1.201   Mean   : 0.4463
##  3rd Qu.: 3.000                   3rd Qu.: 0.000   3rd Qu.: 0.0000
##  Max.   :11.000                   Max.   :37.000   Max.   :37.0000
##  hormomal_contraceptives hormonal_contraceptives_time    iud
##  Min.   :0.0000          Min.   : 0.000               Mode :logical
##  1st Qu.:0.0000          1st Qu.: 0.000               FALSE:775
##  Median :1.0000          Median : 0.250               TRUE :83
##  Mean   :0.5606          Mean   : 1.972
##  3rd Qu.:1.0000          3rd Qu.: 2.000
##  Max.   :1.0000          Max.   :30.000
##     iud_time          std             std_time     condylomatosis
##  Min.   : 0.0000   Mode :logical   Min.   :0.000   Mode :logical
##  1st Qu.: 0.0000   FALSE:779       1st Qu.:0.000   FALSE:814
##  Median : 0.0000   TRUE :79        Median :0.000   TRUE :44
##  Mean   : 0.4446                   Mean   :0.155
##  3rd Qu.: 0.0000                   3rd Qu.:0.000
##  Max.   :19.0000                   Max.   :4.000
##  cervical_condylomatosis vaginal_condylomatosis vulvo_perineal_condylomatosis
##  Min.   :0               Mode :logical          Mode :logical
##  1st Qu.:0               FALSE:854              FALSE:815
##  Median :0               TRUE :4                TRUE :43
##  Mean   :0
##  3rd Qu.:0
##  Max.   :0
##   syphilis       pelvic_inflammatory_disease genital_herpes
##  Mode :logical   Mode :logical               Mode :logical
##  FALSE:840       FALSE:857                   FALSE:857
##  TRUE :18        TRUE :1                     TRUE :1
##
##
##
##  molluscum_contagiosm    aids            hiv          hepatitis_b
##  Mode :logical        Mode :logical   Mode :logical   Mode :logical
##  FALSE:857            FALSE:858       FALSE:840       FALSE:857
##  TRUE :1                              TRUE :18        TRUE :1
##
##
##
##     hpv          diagnosis_num     first_diagnosis_time last_diagnosis_time
##  Mode :logical   Min.   :0.00000   Min.   : 0.0000      Min.   : 0.0000
##  FALSE:856       1st Qu.:0.00000   1st Qu.: 0.0000      1st Qu.: 0.0000
##  TRUE :2         Median :0.00000   Median : 0.0000      Median : 0.0000
##                  Mean   :0.08741   Mean   : 0.5082      Mean   : 0.4814
##                  3rd Qu.:0.00000   3rd Qu.: 0.0000      3rd Qu.: 0.0000
##                  Max.   :3.00000   Max.   :22.0000      Max.   :22.0000
##  dx_cancer         dx_cin          dx_hpv            dx
```

```
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:840       FALSE:849       FALSE:840       FALSE:834
##  TRUE :18        TRUE :9         TRUE :18        TRUE :24
##
##
##
##  hinselmann       schiller        cytology         biopsy
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:823       FALSE:784       FALSE:814       FALSE:803
##  TRUE :35        TRUE :74        TRUE :44        TRUE :55
##
##
##
##  rz_transform_age    rz_transform_sexual_partners
##  Min.   :-3.044808   Min.   :-1.17255
##  1st Qu.:-0.721968   1st Qu.:-0.25666
##  Median :-0.048167   Median :-0.25666
##  Mean   : 0.001286   Mean   : 0.02489
##  3rd Qu.: 0.675406   3rd Qu.: 0.50508
##  Max.   : 3.044808   Max.   : 3.04481
##  rz_transform_first_intercourse rz_transform_pregnancies
##  Min.   :-2.920702              Min.   :-2.33030
##  1st Qu.:-0.744880              1st Qu.:-0.92930
##  Median : 0.140529              Median : 0.01459
##  Mean   : 0.003423              Mean   : 0.02305
##  3rd Qu.: 0.605541              3rd Qu.: 0.70316
##  Max.   : 3.044808              Max.   : 3.04481
##  rz_transform_smoking_time rz_transform_smoke_rate
##  Min.   :-0.1804           Min.   :-0.18044
##  1st Qu.:-0.1804           1st Qu.:-0.18044
##  Median :-0.1804           Median :-0.18044
##  Mean   : 0.0702           Mean   : 0.07021
##  3rd Qu.:-0.1804           3rd Qu.:-0.18044
##  Max.   : 3.0448           Max.   : 3.04481
##  rz_transform_hormonal_contraceptives_time rz_transform_iud
##  Min.   :-0.77211                          Min.   :-0.12140
##  1st Qu.:-0.77211                          1st Qu.:-0.12140
##  Median : 0.03064                          Median :-0.12140
##  Mean   : 0.05307                          Mean   : 0.05051
##  3rd Qu.: 0.60379                          3rd Qu.:-0.12140
##  Max.   : 3.04481                          Max.   : 1.65567
##  rz_transform_std   rz_transform_diagnosis_num
##  Min.   :-0.11552   Min.   :-0.10378
##  1st Qu.:-0.11552   1st Qu.:-0.10378
##  Median :-0.11552   Median :-0.10378
##  Mean   : 0.04972   Mean   : 0.05029
##  3rd Qu.:-0.11552   3rd Qu.:-0.10378
##  Max.   : 1.67911   Max.   : 3.04481
##  rz_transform_first_diagnosis_time rz_transform_last_diagnosis_time
##  Min.   :-0.10378                  Min.   :-0.10378
##  1st Qu.:-0.10378                  1st Qu.:-0.10378
##  Median :-0.10378                  Median :-0.10378
##  Mean   : 0.05602                  Mean   : 0.05601
##  3rd Qu.:-0.10378                  3rd Qu.:-0.10378
##  Max.   : 3.04481                  Max.   : 3.04481
```

```
str(cervicalCA)
```

```
## 'data.frame':    858 obs. of  49 variables:
##  $ X                                     : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ age                                   : int  18 15 34 52 46 42 51 26 45 44 ...
##  $ sexual_partners                       : int  4 1 1 5 3 3 3 1 1 3 ...
##  $ first_intercourse                     : int  15 14 17 16 21 23 17 26 20 15 ...
##  $ pregnancies                           : int  1 1 1 4 4 2 6 3 5 2 ...
##  $ smokes                                : logi  FALSE FALSE FALSE TRUE FALSE FALSE ...
##  $ smoking_time                          : num  0 0 0 37 0 ...
##  $ smoke_rate                            : num  0 0 0 37 0 0 3.4 0 0 2.8 ...
##  $ hormomal_contraceptives               : int  0 0 0 1 1 0 0 1 0 0 ...
##  $ hormonal_contraceptives_time          : num  0 0 0 3 15 0 0 2 0 0 ...
##  $ iud                                   : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ iud_time                              : num  0 0 0 0 0 0 7 7 0 0 ...
##  $ std                                   : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ std_time                              : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ condylomatosis                        : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ cervical_condylomatosis               : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ vaginal_condylomatosis                : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ vulvo_perineal_condylomatosis         : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ syphilis                              : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ pelvic_inflammatory_disease           : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ genital_herpes                        : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ molluscum_contagiosm                  : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ aids                                  : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ hiv                                   : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ hepatitis_b                           : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ hpv                                   : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ diagnosis_num                         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ first_diagnosis_time                  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ last_diagnosis_time                   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ dx_cancer                             : logi  FALSE FALSE FALSE TRUE FALSE FALSE ...
##  $ dx_cin                                : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ dx_hpv                                : logi  FALSE FALSE FALSE TRUE FALSE FALSE ...
##  $ dx                                    : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ hinselmann                            : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ schiller                              : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ cytology                              : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ biopsy                                : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ rz_transform_age                      : num  -1.133 -2.058 0.868 2.429 2.083 ...
##  $ rz_transform_sexual_partners          : num  1.15 -1.173 -1.173 1.606 0.505 ...
##  $ rz_transform_first_intercourse        : num  -0.745 -1.358 0.141 -0.27 1.49 ...
##  $ rz_transform_pregnancies              : num  -0.929 -0.929 -0.929 1.19 1.19 ...
##  $ rz_transform_smoking_time             : num  -0.18 -0.18 -0.18 3.04 -0.18 ...
##  $ rz_transform_smoke_rate               : num  -0.18 -0.18 -0.18 3.04 -0.18 ...
##  $ rz_transform_hormonal_contraceptives_time: num  -0.772 -0.772 -0.772 0.76 2.123 ...
##  $ rz_transform_iud                      : num  -0.121 -0.121 -0.121 -0.121 -0.121 ...
##  $ rz_transform_std                      : num  -0.116 -0.116 -0.116 -0.116 -0.116 ...
##  $ rz_transform_diagnosis_num            : num  -0.104 -0.104 -0.104 -0.104 -0.104 ...
##  $ rz_transform_first_diagnosis_time     : num  -0.104 -0.104 -0.104 -0.104 -0.104 ...
##  $ rz_transform_last_diagnosis_time      : num  -0.104 -0.104 -0.104 -0.104 -0.104 ...
```

```
dim(cervicalCA)
```

```
## [1] 858  49
```

```
class(cervicalCA)
```

```
## [1] "data.frame"
```

```
glimpse(cervicalCA)
```

```
## Rows: 858
## Columns: 49
## $ X                                          <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 1~
## $ age                                        <int> 18, 15, 34, 52, 46, 42, 51, ~
## $ sexual_partners                            <int> 4, 1, 1, 5, 3, 3, 3, 1, 1, 3~
## $ first_intercourse                          <int> 15, 14, 17, 16, 21, 23, 17, ~
## $ pregnancies                                <int> 1, 1, 1, 4, 4, 2, 6, 3, 5, 2~
## $ smokes                                     <lgl> FALSE, FALSE, FALSE, TRUE, F~
## $ smoking_time                               <dbl> 0.000, 0.000, 0.000, 37.000,~
## $ smoke_rate                                 <dbl> 0.0, 0.0, 0.0, 37.0, 0.0, 0.~
## $ hormomal_contraceptives                    <int> 0, 0, 0, 1, 1, 0, 0, 1, 0, 0~
## $ hormonal_contraceptives_time               <dbl> 0.00, 0.00, 0.00, 3.00, 15.0~
## $ iud                                        <lgl> FALSE, FALSE, FALSE, FALSE, ~
## $ iud_time                                   <dbl> 0, 0, 0, 0, 0, 0, 7, 7, 0, 0~
## $ std                                        <lgl> FALSE, FALSE, FALSE, FALSE, ~
## $ std_time                                   <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ condylomatosis                             <lgl> FALSE, FALSE, FALSE, FALSE, ~
## $ cervical_condylomatosis                    <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ vaginal_condylomatosis                     <lgl> FALSE, FALSE, FALSE, FALSE, ~
## $ vulvo_perineal_condylomatosis              <lgl> FALSE, FALSE, FALSE, FALSE, ~
## $ syphilis                                   <lgl> FALSE, FALSE, FALSE, FALSE, ~
## $ pelvic_inflammatory_disease                <lgl> FALSE, FALSE, FALSE, FALSE, ~
## $ genital_herpes                             <lgl> FALSE, FALSE, FALSE, FALSE, ~
## $ molluscum_contagiosm                       <lgl> FALSE, FALSE, FALSE, FALSE, ~
## $ aids                                       <lgl> FALSE, FALSE, FALSE, FALSE, ~
## $ hiv                                        <lgl> FALSE, FALSE, FALSE, FALSE, ~
## $ hepatitis_b                                <lgl> FALSE, FALSE, FALSE, FALSE, ~
## $ hpv                                        <lgl> FALSE, FALSE, FALSE, FALSE, ~
## $ diagnosis_num                              <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ first_diagnosis_time                       <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ last_diagnosis_time                        <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ dx_cancer                                  <lgl> FALSE, FALSE, FALSE, TRUE, F~
## $ dx_cin                                     <lgl> FALSE, FALSE, FALSE, FALSE, ~
## $ dx_hpv                                     <lgl> FALSE, FALSE, FALSE, TRUE, F~
## $ dx                                         <lgl> FALSE, FALSE, FALSE, FALSE, ~
## $ hinselmann                                 <lgl> FALSE, FALSE, FALSE, FALSE, ~
## $ schiller                                   <lgl> FALSE, FALSE, FALSE, FALSE, ~
## $ cytology                                   <lgl> FALSE, FALSE, FALSE, FALSE, ~
## $ biopsy                                     <lgl> FALSE, FALSE, FALSE, FALSE, ~
## $ rz_transform_age                           <dbl> -1.13285321, -2.05809611, 0.~
## $ rz_transform_sexual_partners               <dbl> 1.1496428, -1.1725461, -1.17~
## $ rz_transform_first_intercourse             <dbl> -0.7448795, -1.3575023, 0.14~
## $ rz_transform_pregnancies                   <dbl> -0.9292959, -0.9292959, -0.9~
## $ rz_transform_smoking_time                  <dbl> -0.1804361, -0.1804361, -0.1~
## $ rz_transform_smoke_rate                    <dbl> -0.1804361, -0.1804361, -0.1~
## $ rz_transform_hormonal_contraceptives_time <dbl> -0.77211458, -0.77211458, -0~
## $ rz_transform_iud                           <dbl> -0.1213977, -0.1213977, -0.1~
## $ rz_transform_std                           <dbl> -0.1155205, -0.1155205, -0.1~
## $ rz_transform_diagnosis_num                 <dbl> -0.1037777, -0.1037777, -0.1~
## $ rz_transform_first_diagnosis_time          <dbl> -0.1037777, -0.1037777, -0.1~
## $ rz_transform_last_diagnosis_time           <dbl> -0.1037777, -0.1037777, -0.1~
```

```
head(cervicalCA)
```

```
##   X age sexual_partners first_intercourse pregnancies smokes smoking_time
## 1 1  18               4                15           1  FALSE            0
## 2 2  15               1                14           1  FALSE            0
## 3 3  34               1                17           1  FALSE            0
## 4 4  52               5                16           4   TRUE           37
## 5 5  46               3                21           4  FALSE            0
## 6 6  42               3                23           2  FALSE            0
##   smoke_rate hormomal_contraceptives hormonal_contraceptives_time   iud
## 1          0                       0                            0 FALSE
## 2          0                       0                            0 FALSE
## 3          0                       0                            0 FALSE
## 4         37                       1                            3 FALSE
## 5          0                       1                           15 FALSE
## 6          0                       0                            0 FALSE
##   iud_time   std std_time condylomatosis cervical_condylomatosis
## 1        0 FALSE        0          FALSE                      0
## 2        0 FALSE        0          FALSE                      0
## 3        0 FALSE        0          FALSE                      0
## 4        0 FALSE        0          FALSE                      0
## 5        0 FALSE        0          FALSE                      0
## 6        0 FALSE        0          FALSE                      0
##   vaginal_condylomatosis vulvo_perineal_condylomatosis syphilis
## 1                  FALSE                         FALSE    FALSE
## 2                  FALSE                         FALSE    FALSE
## 3                  FALSE                         FALSE    FALSE
## 4                  FALSE                         FALSE    FALSE
## 5                  FALSE                         FALSE    FALSE
## 6                  FALSE                         FALSE    FALSE
##   pelvic_inflammatory_disease genital_herpes molluscum_contagiosm  aids   hiv
## 1                       FALSE          FALSE                FALSE FALSE FALSE
## 2                       FALSE          FALSE                FALSE FALSE FALSE
## 3                       FALSE          FALSE                FALSE FALSE FALSE
## 4                       FALSE          FALSE                FALSE FALSE FALSE
## 5                       FALSE          FALSE                FALSE FALSE FALSE
## 6                       FALSE          FALSE                FALSE FALSE FALSE
##   hepatitis_b   hpv diagnosis_num first_diagnosis_time last_diagnosis_time
## 1       FALSE FALSE             0                    0                   0
## 2       FALSE FALSE             0                    0                   0
## 3       FALSE FALSE             0                    0                   0
## 4       FALSE FALSE             0                    0                   0
## 5       FALSE FALSE             0                    0                   0
## 6       FALSE FALSE             0                    0                   0
##   dx_cancer dx_cin dx_hpv    dx hinselmann schiller cytology biopsy
## 1     FALSE  FALSE  FALSE FALSE      FALSE    FALSE    FALSE  FALSE
## 2     FALSE  FALSE  FALSE FALSE      FALSE    FALSE    FALSE  FALSE
## 3     FALSE  FALSE  FALSE FALSE      FALSE    FALSE    FALSE  FALSE
## 4      TRUE  FALSE   TRUE FALSE      FALSE    FALSE    FALSE  FALSE
## 5     FALSE  FALSE  FALSE FALSE      FALSE    FALSE    FALSE  FALSE
## 6     FALSE  FALSE  FALSE FALSE      FALSE    FALSE    FALSE  FALSE
##   rz_transform_age rz_transform_sexual_partners rz_transform_first_intercourse
## 1       -1.1328532                    1.149643                    -0.7448795
## 2       -2.0580961                   -1.172546                    -1.3575023
## 3        0.8681143                   -1.172546                     0.1405289
## 4        2.4291594                    1.606039                    -0.2702582
## 5        2.0829847                    0.505078                     1.4900736
## 6        1.7097570                    0.505078                     1.7421048
##   rz_transform_pregnancies rz_transform_smoking_time rz_transform_smoke_rate
```

```
## 1               -0.9292959               -0.1804361              -0.1804361
## 2               -0.9292959               -0.1804361              -0.1804361
## 3               -0.9292959               -0.1804361              -0.1804361
## 4                1.1901358                3.0448080               3.0448080
## 5                1.1901358               -0.1804361              -0.1804361
## 6                0.0145909               -0.1804361              -0.1804361
##   rz_transform_hormonal_contraceptives_time rz_transform_iud rz_transform_std
## 1                                -0.7721146       -0.1213977       -0.1155205
## 2                                -0.7721146       -0.1213977       -0.1155205
## 3                                -0.7721146       -0.1213977       -0.1155205
## 4                                 0.7603734       -0.1213977       -0.1155205
## 5                                 2.1229244       -0.1213977       -0.1155205
## 6                                -0.7721146       -0.1213977       -0.1155205
##   rz_transform_diagnosis_num rz_transform_first_diagnosis_time
## 1                -0.1037777                        -0.1037777
## 2                -0.1037777                        -0.1037777
## 3                -0.1037777                        -0.1037777
## 4                -0.1037777                        -0.1037777
## 5                -0.1037777                        -0.1037777
## 6                -0.1037777                        -0.1037777
##   rz_transform_last_diagnosis_time
## 1                       -0.1037777
## 2                       -0.1037777
## 3                       -0.1037777
## 4                       -0.1037777
## 5                       -0.1037777
## 6                       -0.1037777
```

```
tail(cervicalCA)
```

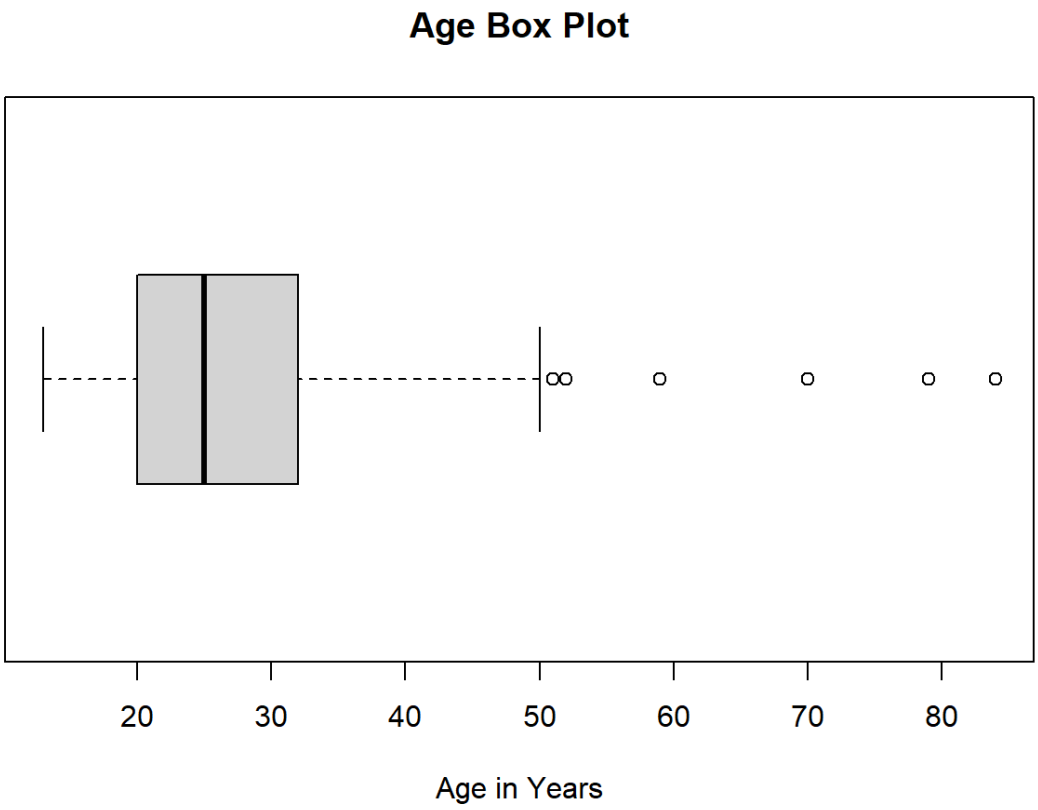
```
##       X age sexual_partners first_intercourse pregnancies smokes smoking_time
## 853 853  43              3                17           3  FALSE            0
## 854 854  34              3                18           0  FALSE            0
## 855 855  32              2                19           1  FALSE            0
## 856 856  25              2                17           0  FALSE            0
## 857 857  33              2                24           2  FALSE            0
## 858 858  29              2                20           1  FALSE            0
##     smoke_rate hormomal_contraceptives hormonal_contraceptives_time   iud
## 853          0                      1                         5.00 FALSE
## 854          0                      0                         0.00 FALSE
## 855          0                      1                         8.00 FALSE
## 856          0                      1                         0.08 FALSE
## 857          0                      1                         0.08 FALSE
## 858          0                      1                         0.50 FALSE
##     iud_time   std std_time condylomatosis cervical_condylomatosis
## 853        0 FALSE        0          FALSE                       0
## 854        0 FALSE        0          FALSE                       0
## 855        0 FALSE        0          FALSE                       0
## 856        0 FALSE        0          FALSE                       0
## 857        0 FALSE        0          FALSE                       0
## 858        0 FALSE        0          FALSE                       0
##     vaginal_condylomatosis vulvo_perineal_condylomatosis syphilis
## 853                  FALSE                         FALSE    FALSE
## 854                  FALSE                         FALSE    FALSE
## 855                  FALSE                         FALSE    FALSE
## 856                  FALSE                         FALSE    FALSE
## 857                  FALSE                         FALSE    FALSE
## 858                  FALSE                         FALSE    FALSE
##     pelvic_inflammatory_disease genital_herpes molluscum_contagiosm  aids   hiv
## 853                       FALSE          FALSE                FALSE FALSE FALSE
## 854                       FALSE          FALSE                FALSE FALSE FALSE
## 855                       FALSE          FALSE                FALSE FALSE FALSE
## 856                       FALSE          FALSE                FALSE FALSE FALSE
## 857                       FALSE          FALSE                FALSE FALSE FALSE
## 858                       FALSE          FALSE                FALSE FALSE FALSE
##     hepatitis_b   hpv diagnosis_num first_diagnosis_time last_diagnosis_time
## 853       FALSE FALSE             0                   0                  0
## 854       FALSE FALSE             0                   0                  0
## 855       FALSE FALSE             0                   0                  0
## 856       FALSE FALSE             0                   0                  0
## 857       FALSE FALSE             0                   0                  0
## 858       FALSE FALSE             0                   0                  0
##     dx_cancer dx_cin dx_hpv    dx hinselmann schiller cytology biopsy
## 853     FALSE  FALSE  FALSE FALSE      FALSE    FALSE    FALSE  FALSE
## 854     FALSE  FALSE  FALSE FALSE      FALSE    FALSE    FALSE  FALSE
## 855     FALSE  FALSE  FALSE FALSE      FALSE    FALSE    FALSE  FALSE
## 856     FALSE  FALSE  FALSE FALSE      FALSE    FALSE     TRUE  FALSE
## 857     FALSE  FALSE  FALSE FALSE      FALSE    FALSE    FALSE  FALSE
## 858     FALSE  FALSE  FALSE FALSE      FALSE    FALSE    FALSE  FALSE
##     rz_transform_age rz_transform_sexual_partners
## 853       1.78349821                   0.5050780
## 854       0.86811427                   0.5050780
## 855       0.67540588                  -0.2566631
## 856      -0.04816689                  -0.2566631
## 857       0.76232294                  -0.2566631
## 858       0.38108094                  -0.2566631
##     rz_transform_first_intercourse rz_transform_pregnancies
```

```
## 853                     0.1405289               0.7031605
## 854                     0.6055411              -2.3302971
## 855                     1.0037832              -0.9292959
## 856                     0.1405289              -2.3302971
## 857                     1.8519853               0.0145909
## 858                     1.2742887              -0.9292959
##     rz_transform_smoking_time rz_transform_smoke_rate
## 853                -0.1804361             -0.1804361
## 854                -0.1804361             -0.1804361
## 855                -0.1804361             -0.1804361
## 856                -0.1804361             -0.1804361
## 857                -0.1804361             -0.1804361
## 858                -0.1804361             -0.1804361
##     rz_transform_hormonal_contraceptives_time rz_transform_iud rz_transform_std
## 853                                1.0232917       -0.1213977       -0.1155205
## 854                               -0.7721146       -0.1213977       -0.1155205
## 855                                1.4812755       -0.1213977       -0.1155205
## 856                               -0.1155205       -0.1213977       -0.1155205
## 857                               -0.1155205       -0.1213977       -0.1155205
## 858                                0.1804361       -0.1213977       -0.1155205
##     rz_transform_diagnosis_num rz_transform_first_diagnosis_time
## 853                -0.1037777                         -0.1037777
## 854                -0.1037777                         -0.1037777
## 855                -0.1037777                         -0.1037777
## 856                -0.1037777                         -0.1037777
## 857                -0.1037777                         -0.1037777
## 858                -0.1037777                         -0.1037777
##     rz_transform_last_diagnosis_time
## 853                       -0.1037777
## 854                       -0.1037777
## 855                       -0.1037777
## 856                       -0.1037777
## 857                       -0.1037777
## 858                       -0.1037777
```

Data Analysis: Box Plots

I create a few box plots to analyze the measures of spread for a few variables. I chose to work with age, sexual partners, first intercourse, and pregnancies. These variables are numerical, as opposed to most of the other variables in the data set that are Boolean. Age is important because it gives the data scientist an idea of the ages associated with cervical cancer symptoms and risks. By analyzing the number of sexual partners, doctors can assess the importance of sexual partners as a risk factor for cervical cancer and other similar diseases. First intercourse data gives doctors an idea of how early on women are exposed to HPV. This gives them a time frame to work with when planning prevention. Pregnancy data helps data scientists further analyze the spread of HPV. Knowing the rate of perinatal transmission, data scientists can further examine the spread of this disease as it persists in our population.

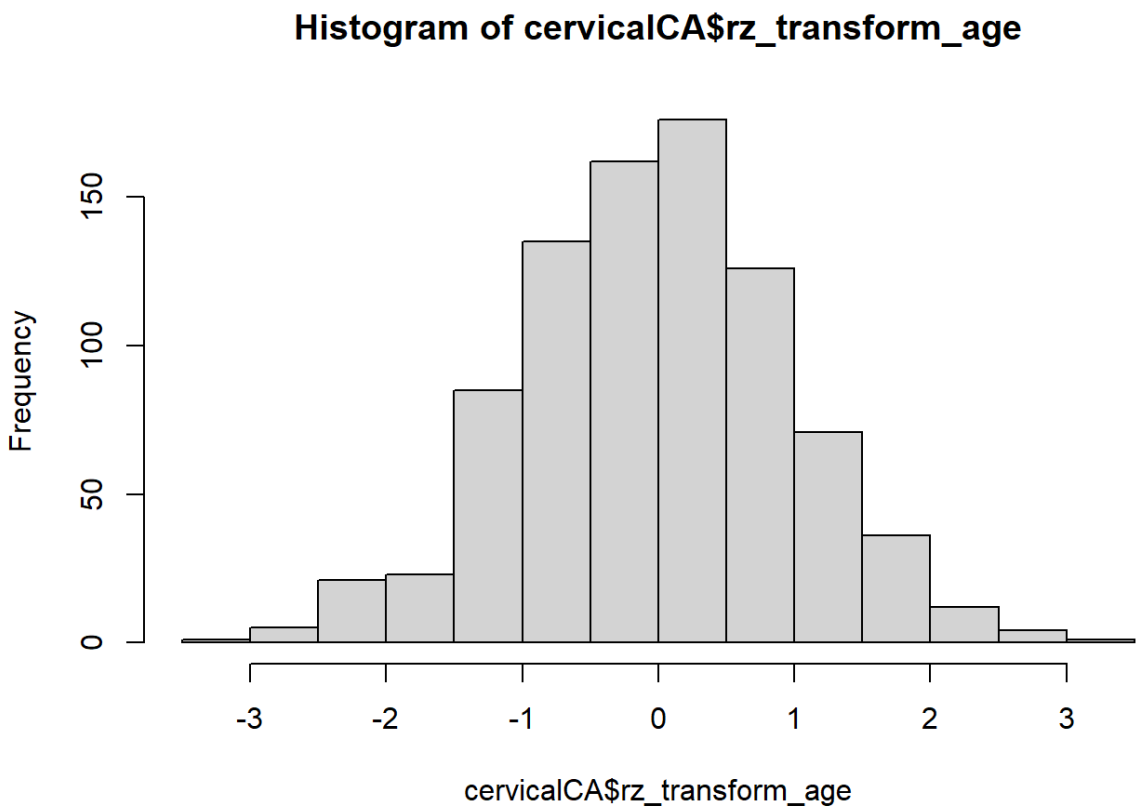
```
#
# Basic boxplots
boxplot(cervicalCA$age, horizontal = TRUE, main = "Age Box Plot", xlab = "Age in Years")
```



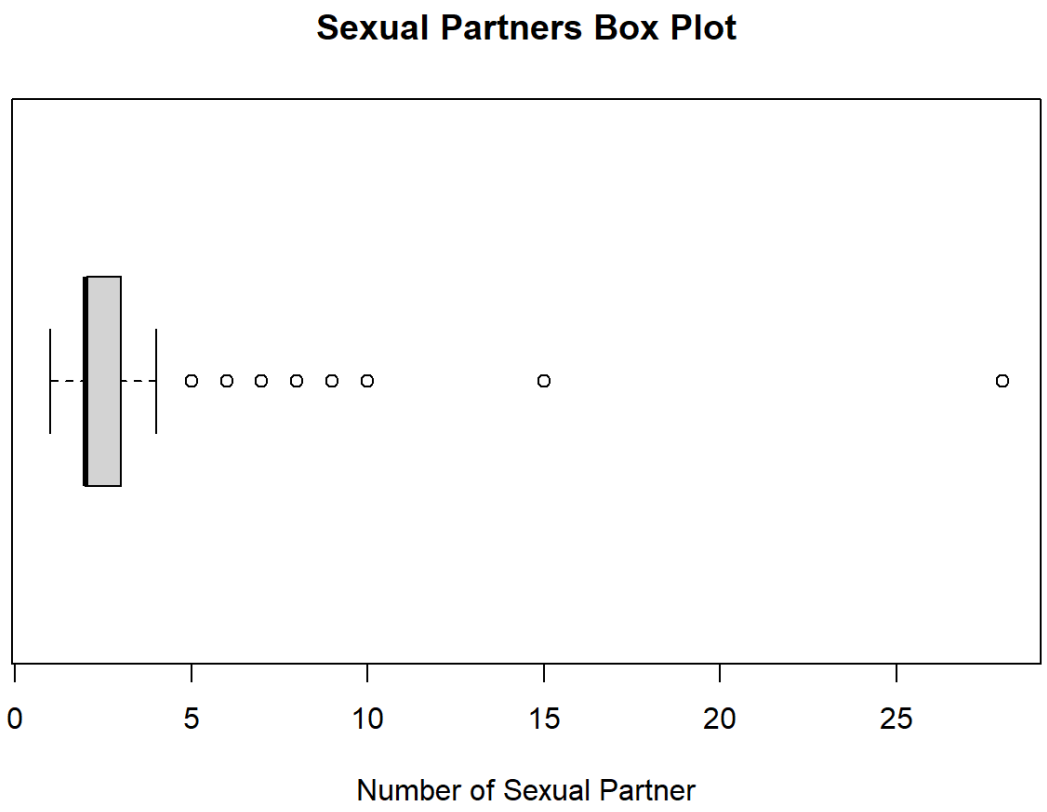

```
quantile(cervicalCA$age)
```

```
##   0%  25%  50%  75% 100%
##   13   20   25   32   84
```

```
hist(cervicalCA$rz_transform_age)
```



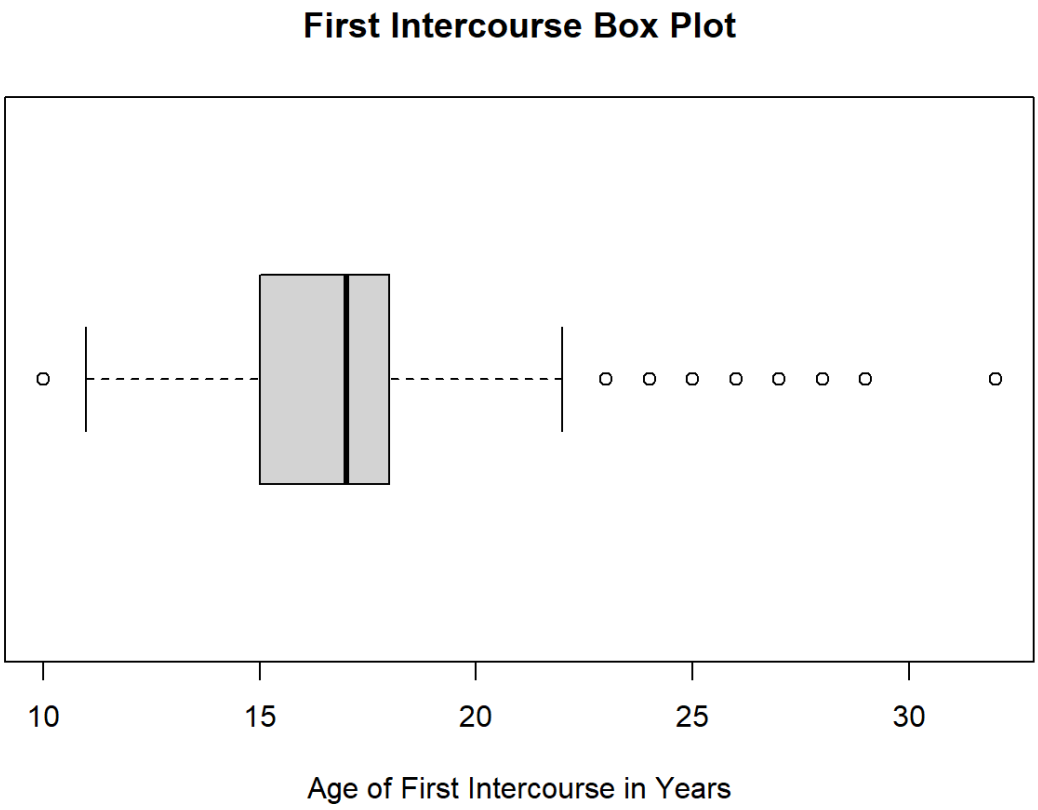

```
# This is included because it tells us that most of
# the women that were studied were relatively young
# with some older women as outliers.
#
boxplot(cervicalCA$sexual_partners, horizontal = TRUE, main = "Sexual Partners Box Plot", xlab = "Number
of Sexual Partner")
```




```
quantile(cervicalCA$sexual_partners)
```

```
##   0%  25%  50%  75% 100%
##    1    2    2    3   28
```

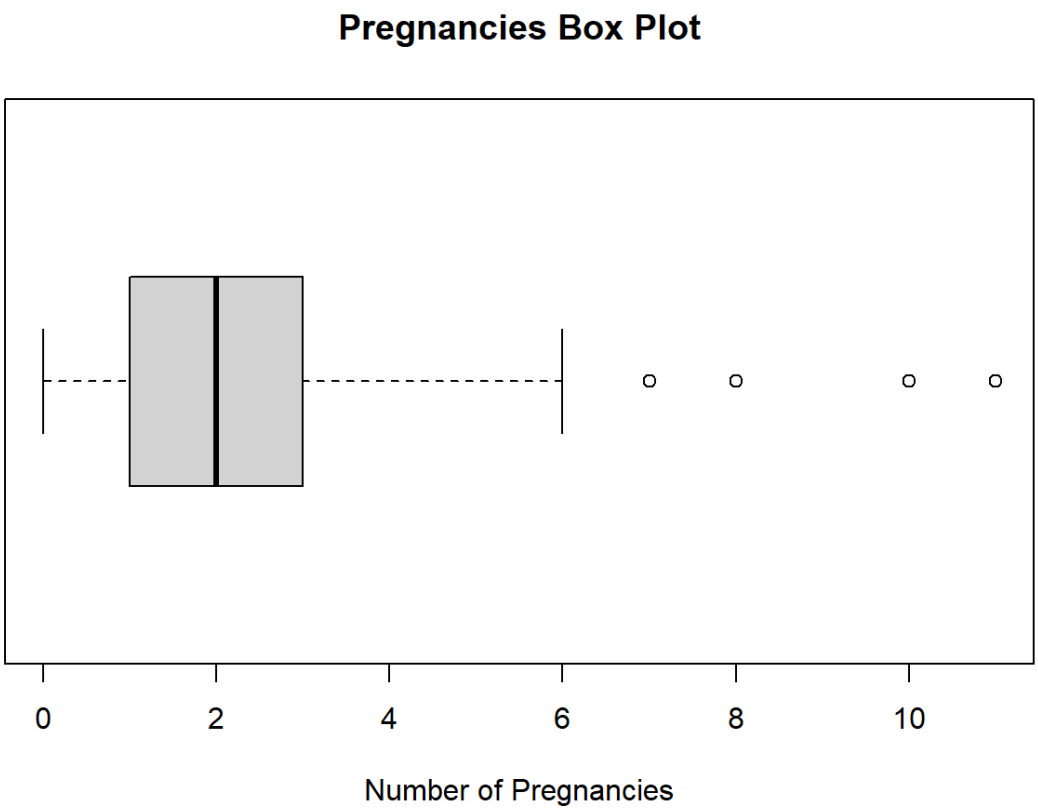
```
#
# This boxplot is included because it tells us that
# most of the women had a few sexual partners, with a
# few outliers with many sexual partners. Since the
# main cause of Cervical cancer is HPV (an STD), it
# only makes sense that more sexual partners increases
# the likelihood of transmission.
#
boxplot(cervicalCA$first_intercourse, horizontal = TRUE, main = "First Intercourse Box Plot", xlab = "Ag
e of First Intercourse in Years")
```




```
quantile(cervicalCA$first_intercourse)
```

```
##   0%  25%  50%  75% 100%
##   10   15   17   18   32
```

```
#
# This box plot is included because it tells us that
# most women in this dataset became sexually active
# relatively early (shortly after puberty),
# exposing them to stds early on in their lives.
#
boxplot(cervicalCA$pregnancies, horizontal = TRUE, main = "Pregnancies Box Plot", xlab = "Number of Preg
nancies")
```




```
quantile(cervicalCA$pregnancies)
```

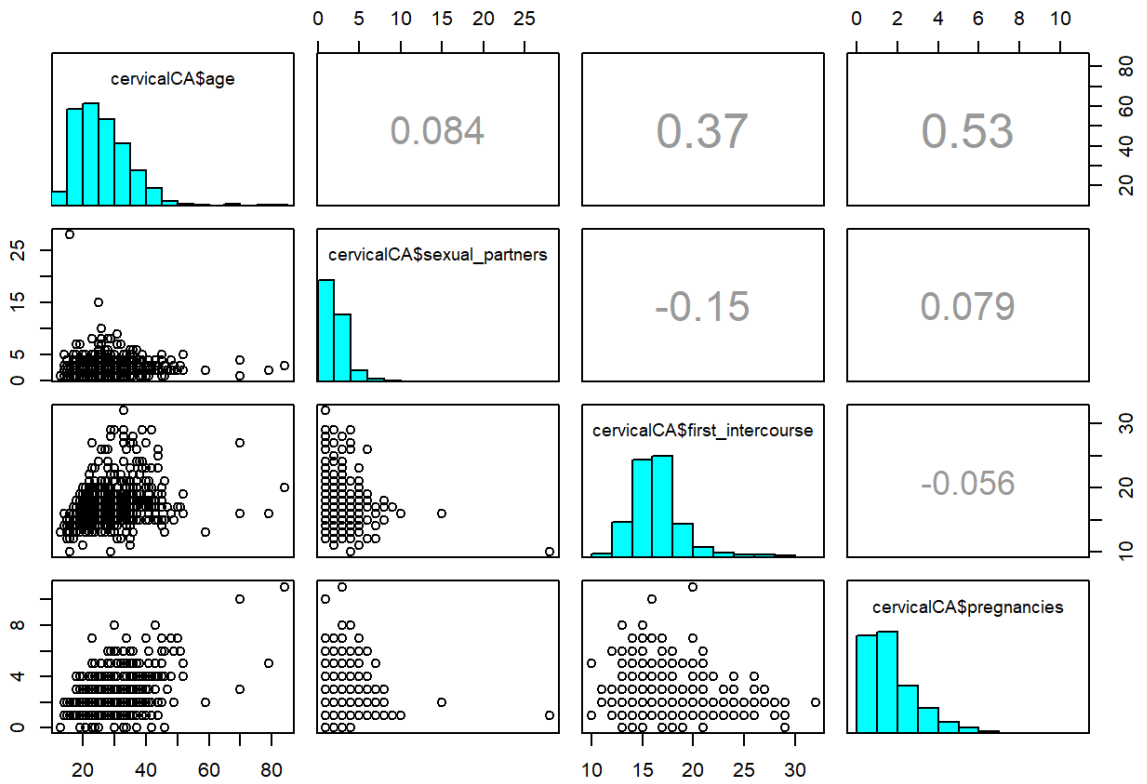
```
##   0%  25%  50%  75% 100%
##    0    1    2    3   11
```

```
#
# This box plot is included because it gives us an
# idea of the perinatal spread of HPV
```

Pairwise Correlation Plot

The pairwise correlation plot gives a brief overview of possible correlation between different variables. Here I look at age, sexual partners, first intercourse, and pregnancies to see if I can extrapolate any correlative relationships.

```
#
# Pairwise correlation plot
pairs(~cervicalCA$age + cervicalCA$sexual_partners + cervicalCA$first_intercourse + cervicalCA$pregnanci
es, upper.panel = panel.cor, diag.panel = panel.hist)
```

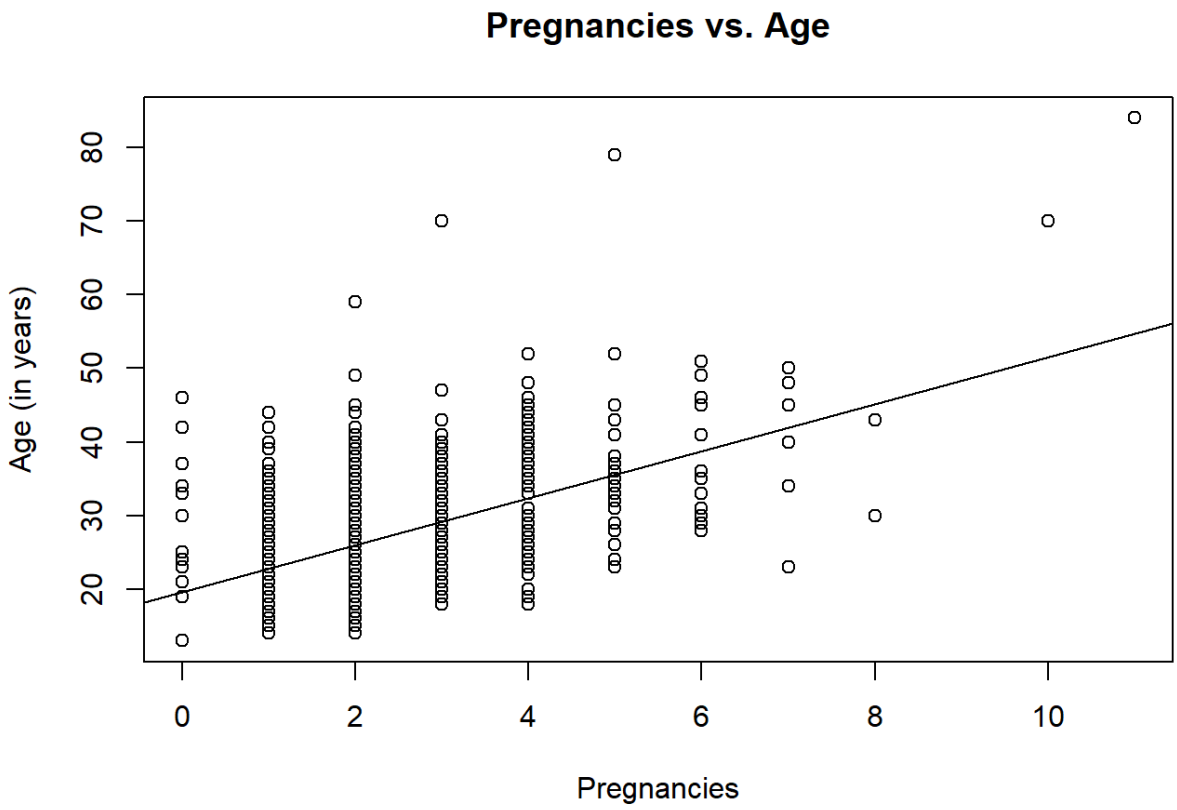



```
#
# This plot shows possible correlation between the
# variables age, sexual partners, first intercourse,
# and pregnancies. The correlation coefficients
# a moderate positive linear relationship between age
# and pregnancies. The coefficients also suggest a
# weak positive linear relationship between age and
# first sexual intercourse. The other coefficients
# don't show defensible relationships.
```

Linear Regression

Here I perform a linear regression of age and pregnancies. Based on the plot, pregnancies tend to increase with age, showing that many women may be passing HPV to their children. Also, the fact that mostly HPV can stay dormant for years supports the fact that women may unknowingly spread the disease to their children.

```
#
# Linear regression of age and pregnancies
plot(cervicalCA$age~cervicalCA$pregnancies,
     main="Pregnancies vs. Age",
     xlab="Pregnancies", ylab="Age (in years)")
fitline <- lm(cervicalCA$age~cervicalCA$pregnancies)
abline(fitline)
```




```
summary(fitline)
```

```
##
## Call:
## lm(formula = cervicalCA$age ~ cervicalCA$pregnancies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.948  -4.999  -1.474   4.001  43.431
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)              19.6191     0.4684   41.88   <2e-16 ***
## cervicalCA$pregnancies   3.1899     0.1763   18.09   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.232 on 856 degrees of freedom
## Multiple R-squared:  0.2766, Adjusted R-squared:  0.2757
## F-statistic: 327.2 on 1 and 856 DF,  p-value: < 2.2e-16
```

Bayesian Information Criterion (BIC) Analysis

BIC analysis allows me to determine causation between any variables. I use it here to check for causation between age and hormonal contraceptives time. I first measure the BIC score of the target to itself, then measure it against age. If the second score is 10 less than the first score, than there is likely to be a causal relationship between the two variables. Here I determine that age likely causes people to use hormonal contraceptives for a longer time.

```
#
# BIC modeling for age -> hormonal contraceptives time
BIC(lm(cervicalCA$hormonal_contraceptives_time~1)) # 4644
```

```
## [1] 4644.483
```

```
BIC(lm(cervicalCA$hormonal_contraceptives_time~cervicalCA$age)) # 4570
```

```
## [1] 4570.944
```

```
# Causal
```

Separating Data

Here I separate the data based on a low number of sexual partners to examine patterns within just this group of people. I compare the BIC scores of condylomatosis and the diagnosis number and determine that it is causal. I then compare the BIC scores of HIV and the diagnosis number. Though this relationship is causal, it isn't as strong as the previous model because the BIC score difference is less. This makes sense because condylomatosis is much easier to diagnose than HIV, especially among people with low sexual parters (knowing that having low sexual partners decreases the risk of STD transmission).

```
#
# Separating based on a low number of sexual partners
lowSexualPartners <- filter(cervicalCA, sexual_partners <= median(cervicalCA$sexual_partners))
#
# BIC modeling for condylomatosis -> diagnosis num for those
# with low sexual partners
BIC(lm(lowSexualPartners$diagnosis_num~1)) # 248
```

```
## [1] 247.9661
```

```
BIC(lm(lowSexualPartners$diagnosis_num~lowSexualPartners$condylomatosis)) # -149
```

```
## [1] -149.4535
```

```
# Causal
#
# BIC modeling for hiv -> diagnosis num for those
# with low sexual partners
BIC(lm(lowSexualPartners$diagnosis_num~1)) # 248
```

```
## [1] 247.9661
```

```
BIC(lm(lowSexualPartners$diagnosis_num~lowSexualPartners$hiv)) # 32
```

```
## [1] 31.88176
```

```
#
# Causal, but less than condylomatosis which hints at
# less recognition of hiv than condylomatosis. This
# makes sense because the group of people have less
# sexual partners, making diagnosis less common.
```

More Separating Data

Here I separate the data based on whether the person has an STD. Again, I do this to analyze patters within this subset of people. Here I'm analyzing the causal relationship between HPV and a change in cancer. Based on the decrease BIC score difference, I determined that there is a strong causal relationship between these two variables. This makes sense since almost all of those who have cervical cancer developed their cancer from an HPV infection.

```
#
# Separating based on having an std
peopleWithSTDs <- filter(cervicalCA, std == TRUE)
#
# BIC modeling for hpv -> dx_cancer
# for people with stds
BIC(lm(peopleWithSTDs$dx_cancer~1)) # -59
```

```
## [1] -59.52232
```

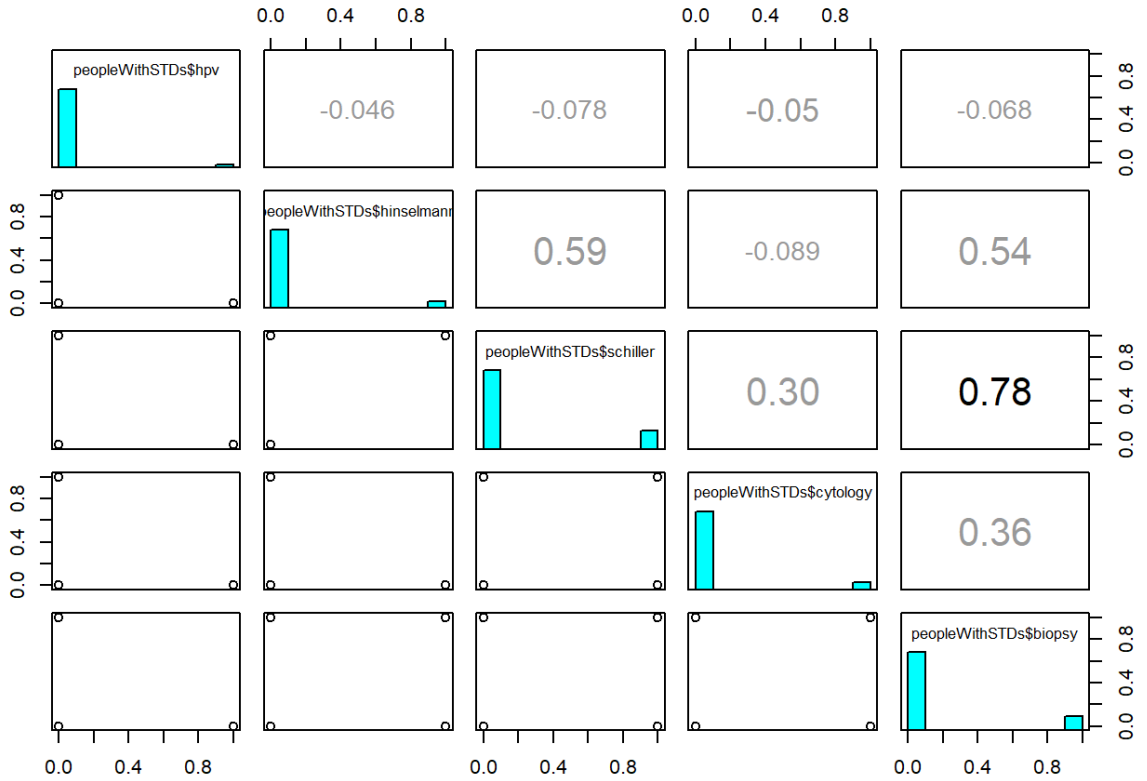
```
BIC(lm(peopleWithSTDs$dx_cancer~peopleWithSTDs$hpv)) # -5638
```

```
## [1] -5637.787
```

```
# Causal
```

To assess the danger of detecting HPV I made this pairwise plot to measure the correlation between HPV and the labs done to detect cervical cancer. This plot shows that HPV is generally not correlated with the cervical cancer tests, making it difficult for doctors to detect HPV when trying to diagnose cervical cancer.

```
#
# Pairwise plot of test results from people with stds with hpv.
pairs(~peopleWithSTDs$hpv + peopleWithSTDs$hinselmann + peopleWithSTDs$schiller + peopleWithSTDs$cytolog
y + peopleWithSTDs$biopsy, upper.panel = panel.cor, diag.panel = panel.hist)
```




```
# This shows that hpv generally goes unnoticed by these tests.
# Also the tests are correlated with each other (with the
# exception of cytology and hinselmann).
```

To further prove a lack of correlation between HPV and the cervical cancer tests, I designed a multiple linear model which produces a P value. This P value is used to assess the potential for a correlative relationship. The calculated P value was 0.97 which further proves that the tests were not statistically significant enough to support a correlative relationship

```
# Multifit of the tests based on hpv
multifit <- lm(peopleWithSTDs$hpv ~ peopleWithSTDs$hinselmann + peopleWithSTDs$schiller + peopleWithSTDs
$cytology + peopleWithSTDs$biopsy)
summary (multifit)
```

```
##
## Call:
## lm(formula = peopleWithSTDs$hpv ~ peopleWithSTDs$hinselmann +
##     peopleWithSTDs$schiller + peopleWithSTDs$cytology + peopleWithSTDs$biopsy)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.03212 -0.03212 -0.03212 -0.02209  0.96788
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   0.032123   0.020498   1.567    0.121
## peopleWithSTDs$hinselmannTRUE -0.007599   0.094165  -0.081    0.936
## peopleWithSTDs$schillerTRUE   -0.022731   0.079989  -0.284    0.777
## peopleWithSTDs$cytologyTRUE   -0.017904   0.075019  -0.239    0.812
## peopleWithSTDs$biopsyTRUE     -0.002152   0.086551  -0.025    0.980
##
## Residual standard error: 0.1617 on 74 degrees of freedom
## Multiple R-squared:  0.006983,   Adjusted R-squared:  -0.04669
## F-statistic: 0.1301 on 4 and 74 DF,  p-value: 0.971
```

```
#
# The P value of 0.97 shows that the
# tests are not statistically significant enough
# to prove correlation.
#
# EOF (End of File)
```

Work Cited

“Basic Information about Cervical Cancer.” Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, 12 Jan. 2021, www.cdc.gov/cancer/cervical/basic_info/index.htm.

“Cervical Biopsy.” Johns Hopkins Medicine, John Hopkins Medicine, 2021, www.hopkinsmedicine.org/health/treatment-tests-and-therapies/cervical-biopsy.

“Cervical Cancer - Symptoms and Signs.” Cancer.Net, 26 Feb. 2021, www.cancer.net/cancer-types/cervical-cancer/symptoms-and-signs.

“Cervical Cancer Statistics: Key Facts about Cervical Cancer.” American Cancer Society, 2021, www.cancer.org/cancer/cervical-cancer/about/key-statistics.html.

“Cervical Cancer.” Mayo Clinic, Mayo Foundation for Medical Education and Research, 17 June 2021, www.mayoclinic.org/diseases-conditions/cervical-cancer/symptoms-causes/syc-20352501.

Ih-Acr. “Patient Education Library.” The Angeles Clinic, The Angeles Clinic, 2021, www.theangelesclinic.org/Home/ResearchClinicalTrials/PatientEducationLibrary/tabid/19283/ctl/View/mid/35200/Default.aspx?ContentPubID=30.