Journey through the Exploration of Cervical Data

Samuel Dummer

9/21/2021

Abstract

Cervical cancer is a cancer in the cervix, this means it only affects women. The survival rate of this cancer is not super high which means that there are many women every year that die from this cancer. In fact, it is actually the fourth most deadly cancer for women. This specific dataset contains information about 858 different patients. There are also many different variables that define these patients such as age, number of sexual partners, etc.. When exploring this data, there were many different correlations found. Many of them were fairly weak, but they were still present. These correlations include: smoking, number of sexual partners, and HPV. The only causation found was that HPV has a high chance of causing women to develop cervical cancer. There were many other STDs that were tested for in these patients, but it didn't seem like any of them came as close as HPV did in terms of correlation with cervical cancer.

Introduction

As stated above, cervical cancer is a cancer that only affects women. It specifically occurs when the cells in a women's cervix start to mutate. The cervix is the part that connects the uterus and the vagina. The leading cause of this cancer in women is HPV. Contracting HPV does not mean that you will develop cervical cancer, but 90% of all patients with cervical cancer also had HPV. This is because in some women, HPV can survive many years after it has been treated and through these years can help cause the cervix cells to mutate into cancerous cells. Not only is this cancer also the fourth most deadly cancer, with a 68% mortality rate, but it is also the fourth most common cancer in women worldwide.

For a long time, cervical cancer will cause no symptoms and stay fairly dormant, but once it starts to become active, there are several symptoms it may cause. These include, vaginal bleeding outside of menstrual cycles, pain during intercourse, or watery vaginal discharge with a strong smell.

The best way to treat cervical cancer is actually to prevent it before it happens. First of all, getting the HPV vaccine since there is high causation between HPV and cervical cancer. Next, you can regularly test since many tests detect any cancerous cells whether or not there are symptoms or not. Preventing the spread of STDs through safe sex also plays a key role in preventing cervical cancer. Lastly, not smoking is also helpful since there has been some correlation found between smoking and developing cervical cancer. Smoking can cause your immune system to become weaker which provides a higher possibility for the development of cervical cancer.

In this study, we took data describing 858 patients and performed some cleaning and data munging on it. This was done by changing some column names, converting some numerical values to logical values since that was the for some of the data was supposed to be in for some columns. We also replaced many "NA" values with either the mean or "FALSE" depending on whether it was a numerical values or a logical values. Next, we performed some basic exploratory data analysis (EDA) on this cleaned data. This EDA helped us find any correlation or causation between the many variables in the data set. This finding correlation was done in many ways. To start off, we created a pairwise correlation plot and didn't find any new knowledge or discoveries. Then we created some new data sets that were grouped by whether the patient had cancer or not and then created some summary charts of different variables such as smoking, age, and age when they had their first sexual intercourse. Then, we performed a multiple linear regression to test to find anything interesting, but there didn't seem to be any correlation between all the data tested. Lastly, we performed various Bayesian Information Criterion (BIC) models to try to find any causation. The only causation found through this method was between HPV and developing cervical cancer.

Setting Up Directories, Cleaning Environment, etc.

Before we can perform any sort of EDA, we need to set up the working directory and clean the environment. This helps to clear up the the environment from the previous time the environment was used. Next we also need to load in any libraries that we believe will be used in our code. In this specific script, I loaded in the "tidyverse" and "stringr" libraries. These libraries help to give us many different/more efficient functions to use. Lastly, we also loaded in a user function in case we needed to use some functions from there.

```
rm(list=ls())
setwd("C:/Users/isabe/Desktop/RFLoder")
source("myfunctions.R")
library(stringr)
library(tidyverse)
```

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

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.3     v dplyr   1.0.7
## v tidyr   1.1.3     v forcats 0.5.1
## v readr   2.0.1

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

Loading in Cleaned Dataset

Since we have already cleaned the data set in a previous R Script and then we saved it to a new .csv file all we need to do is load in the new files. We do so by using the read_csv command and reading in r file named "ccdataMod.csv".

```
cervical <- read_csv("ccdataMod.csv")
```

```
## New names:
## * `` -> ...1

## Rows: 858 Columns: 39

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (15): ...1, age, number_of_sexual_partners, first_sexual_intercourse, nu...
## lgl (24): smokes, hormonal_contraceptives, iud, stds, stds_condylomatosis, s...

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Summarizing/Looking at the Structure of the Data

Now that we have read in the dataset we can start to look over the structure and summary of the data. This part is fairly quick but very crucial it helps to tell us what the general makeup of this data set. There were many different ways that we looked at the data, we started with the names() function to give us a quick rundown of all the variable names. Then, we used the dim() function to give us the number of variables and number of observations. Next, we used the str() function to give us the general structure of the data. This helps to give us the type of variable that each variable is. Then, the glimpse() function was used which gives similar data as the str() function, but just in a different format. Then head() and tail() are used to show us the beginning and end of the data which helps us see if there is any problems when reading in the data. Lastly, we used the summary() function which gives us a summary of each column. This includes the mean, median, max, and min for numerical values and the number of trues and falses for logical data. Overall, using these functions helps to give us a good idea of the makeup of the data.

```
names(cervical)
```

```
##  [1] "...1"                               "age"
##  [3] "number_of_sexual_partners"          "first_sexual_intercourse"
##  [5] "num_of_pregnancies"                 "smokes"
##  [7] "smokes_years"                       "smokes_packs_year"
##  [9] "hormonal_contraceptives"            "hormonal_contraceptives_years"
## [11] "iud"                                "iud_years"
## [13] "stds"                               "stds_number"
## [15] "stds_condylomatosis"                "stds_cervical_condylomatosis"
## [17] "stds_vaginal_condylomatosis"        "stds_vulvo_perineal_condylomatosis"
## [19] "stds_syphilis"                      "stds_pelvic_inflammatory_disease"
## [21] "stds_genital_herpes"                "stds_molluscum_contagiosum"
## [23] "stds_aids"                          "stds_hiv"
## [25] "stds_hepatitis_b"                   "stds_hpv"
## [27] "stds_number_of_diagnosis"           "dx_cancer"
## [29] "dx_cin"                             "dx_hpv"
## [31] "dx"                                 "hinselmann"
## [33] "schiller"                           "citology"
## [35] "biopsy"                             "age.rz"
## [37] "pregnancy.rz"                       "sexualpartners.rz"
## [39] "hormonal.rz"
```

```
dim(cervical)
```

```
## [1] 858  39
```

```
str(cervical)
```

```
## spec_tbl_df [858 x 39] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ...1                              : num [1:858] 1 2 3 4 5 6 7 8 9 10 ...
##  $ age                               : num [1:858] 18 15 34 52 46 42 51 26 45 44 ...
##  $ number_of_sexual_partners         : num [1:858] 4 1 1 5 3 3 3 1 1 3 ...
##  $ first_sexual_intercourse          : num [1:858] 15 14 16 16 21 23 17 26 20 15 ...
##  $ num_of_pregnancies                : num [1:858] 1 1 1 4 4 2 6 3 5 2 ...
##  $ smokes                            : logi [1:858] FALSE FALSE FALSE TRUE FALSE FALSE ...
##  $ smokes_years                      : num [1:858] 0 0 0 37 0 ...
##  $ smokes_packs_year                 : num [1:858] 0 0 0 37 0 0 3.4 0 0 2.8 ...
##  $ hormonal_contraceptives           : logi [1:858] FALSE FALSE FALSE TRUE TRUE FALSE ...
##  $ hormonal_contraceptives_years     : num [1:858] 0 0 0 3 15 0 0 2 0 0 ...
##  $ iud                               : logi [1:858] FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ iud_years                         : num [1:858] 0 0 0 0 0 0 7 7 0 0 ...
##  $ stds                              : logi [1:858] FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ stds_number                       : num [1:858] 0 0 0 0 0 0 0 0 0 0 ...
##  $ stds_condylomatosis               : logi [1:858] FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ stds_cervical_condylomatosis      : logi [1:858] FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ stds_vaginal_condylomatosis       : logi [1:858] FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ stds_vulvo_perineal_condylomatosis: logi [1:858] FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ stds_syphilis                     : logi [1:858] FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ stds_pelvic_inflammatory_disease  : logi [1:858] FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ stds_genital_herpes               : logi [1:858] FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ stds_molluscum_contagiosum        : logi [1:858] FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ stds_aids                         : logi [1:858] FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ stds_hiv                          : logi [1:858] FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ stds_hepatitis_b                  : logi [1:858] FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ stds_hpv                          : logi [1:858] FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ stds_number_of_diagnosis          : num [1:858] 0 0 0 0 0 0 0 0 0 0 ...
##  $ dx_cancer                         : logi [1:858] FALSE FALSE FALSE TRUE FALSE FALSE ...
##  $ dx_cin                            : logi [1:858] FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ dx_hpv                            : logi [1:858] FALSE FALSE FALSE TRUE FALSE FALSE ...
##  $ dx                                : logi [1:858] FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ hinselmann                        : logi [1:858] FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ schiller                          : logi [1:858] FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ citology                          : logi [1:858] FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ biopsy                            : logi [1:858] FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ age.rz                            : num [1:858] -1.133 -2.058 0.868 2.429 2.083 ...
##  $ pregnancy.rz                      : num [1:858] -0.929 -0.929 -0.929 1.19 1.19 ...
##  $ sexualpartners.rz                 : num [1:858] 1.15 -1.173 -1.173 1.606 0.549 ...
##  $ hormonal.rz                       : num [1:858] -1.01 -1.01 -1.01 0.76 2.12 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   ...1 = col_double(),
##   ..   age = col_double(),
##   ..   number_of_sexual_partners = col_double(),
##   ..   first_sexual_intercourse = col_double(),
##   ..   num_of_pregnancies = col_double(),
##   ..   smokes = col_logical(),
##   ..   smokes_years = col_double(),
##   ..   smokes_packs_year = col_double(),
##   ..   hormonal_contraceptives = col_logical(),
##   ..   hormonal_contraceptives_years = col_double(),
##   ..   iud = col_logical(),
##   ..   iud_years = col_double(),
##   ..   stds = col_logical(),
##   ..   stds_number = col_double(),
##   ..   stds_condylomatosis = col_logical(),
##   ..   stds_cervical_condylomatosis = col_logical(),
##   ..   stds_vaginal_condylomatosis = col_logical(),
##   ..   stds_vulvo_perineal_condylomatosis = col_logical(),
##   ..   stds_syphilis = col_logical(),
##   ..   stds_pelvic_inflammatory_disease = col_logical(),
##   ..   stds_genital_herpes = col_logical(),
##   ..   stds_molluscum_contagiosum = col_logical(),
##   ..   stds_aids = col_logical(),
##   ..   stds_hiv = col_logical(),
##   ..   stds_hepatitis_b = col_logical(),
##   ..   stds_hpv = col_logical(),
##   ..   stds_number_of_diagnosis = col_double(),
##   ..   dx_cancer = col_logical(),
##   ..   dx_cin = col_logical(),
##   ..   dx_hpv = col_logical(),
##   ..   dx = col_logical(),
##   ..   hinselmann = col_logical(),
##   ..   schiller = col_logical(),
##   ..   citology = col_logical(),
##   ..   biopsy = col_logical(),
##   ..   age.rz = col_double(),
##   ..   pregnancy.rz = col_double(),
##   ..   sexualpartners.rz = col_double(),
##   ..   hormonal.rz = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
glimpse(cervical)
```

```
## Rows: 858
## Columns: 39
## $ ...1                               <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, ~
## $ age                                <dbl> 18, 15, 34, 52, 46, 42, 51, 26, 45,~
## $ number_of_sexual_partners          <dbl> 4, 1, 1, 5, 3, 3, 3, 1, 1, 3, 3, 1,~
## $ first_sexual_intercourse           <dbl> 15, 14, 16, 16, 21, 23, 17, 26, 20,~
## $ num_of_pregnancies                 <dbl> 1, 1, 1, 4, 4, 2, 6, 3, 5, 2, 4, 3,~
## $ smokes                             <lgl> FALSE, FALSE, FALSE, TRUE, FALSE, F~
## $ 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            <lgl> FALSE, FALSE, FALSE, TRUE, TRUE, FA~
## $ hormonal_contraceptives_years      <dbl> 0.00, 0.00, 0.00, 3.00, 15.00, 0.00~
## $ iud                                <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, ~
## $ iud_years                          <dbl> 0, 0, 0, 0, 0, 0, 7, 7, 0, 0, 0, 0,~
## $ stds                               <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, ~
## $ stds_number                        <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ stds_condylomatosis                <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, ~
## $ stds_cervical_condylomatosis       <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, ~
## $ stds_vaginal_condylomatosis        <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, ~
## $ stds_vulvo_perineal_condylomatosis <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, ~
## $ stds_syphilis                      <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, ~
## $ stds_pelvic_inflammatory_disease   <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, ~
## $ stds_genital_herpes                <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, ~
## $ stds_molluscum_contagiosum         <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, ~
## $ stds_aids                          <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, ~
## $ stds_hiv                           <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, ~
## $ stds_hepatitis_b                   <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, ~
## $ stds_hpv                           <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, ~
## $ stds_number_of_diagnosis           <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ dx_cancer                          <lgl> FALSE, FALSE, FALSE, TRUE, FALSE, F~
## $ dx_cin                             <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, ~
## $ dx_hpv                             <lgl> FALSE, FALSE, FALSE, TRUE, FALSE, F~
## $ dx                                 <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, ~
## $ hinselmann                         <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, ~
## $ schiller                           <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, ~
## $ citology                           <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, ~
## $ biopsy                             <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, ~
## $ age.rz                             <dbl> -1.13285321, -2.05809611, 0.8681142~
## $ pregnancy.rz                       <dbl> -0.9292959, -0.9292959, -0.9292959,~
## $ sexualpartners.rz                  <dbl> 1.1496428, -1.1725461, -1.1725461, ~
## $ hormonal.rz                        <dbl> -1.0062008, -1.0062008, -1.0062008,~
```

```
head(cervical)
```

```
## # A tibble: 6 x 39
##    ...1   age number_of_sexual_par~ first_sexual_interc~ num_of_pregnanc~ smokes
##   <dbl> <dbl>                <dbl>                <dbl>            <dbl> <lgl>
## 1     1    18                    4                   15                1 FALSE
## 2     2    15                    1                   14                1 FALSE
## 3     3    34                    1                   16                1 FALSE
## 4     4    52                    5                   16                4 TRUE
## 5     5    46                    3                   21                4 FALSE
## 6     6    42                    3                   23                2 FALSE
## # ... with 33 more variables: smokes_years <dbl>, smokes_packs_year <dbl>,
## #   hormonal_contraceptives <lgl>, hormonal_contraceptives_years <dbl>,
## #   iud <lgl>, iud_years <dbl>, stds <lgl>, stds_number <dbl>,
## #   stds_condylomatosis <lgl>, stds_cervical_condylomatosis <lgl>,
## #   stds_vaginal_condylomatosis <lgl>,
## #   stds_vulvo_perineal_condylomatosis <lgl>, stds_syphilis <lgl>,
## #   stds_pelvic_inflammatory_disease <lgl>, stds_genital_herpes <lgl>, ...
```

```
tail(cervical)
```

```
## # A tibble: 6 x 39
##    ...1   age number_of_sexual_par~ first_sexual_interc~ num_of_pregnanc~ smokes
##   <dbl> <dbl>                <dbl>                <dbl>            <dbl> <lgl>
## 1   853    43                    3                   17                3 FALSE
## 2   854    34                    3                   18                0 FALSE
## 3   855    32                    2                   19                1 FALSE
## 4   856    25                    2                   17                0 FALSE
## 5   857    33                    2                   24                2 FALSE
## 6   858    29                    2                   20                1 FALSE
## # ... with 33 more variables: smokes_years <dbl>, smokes_packs_year <dbl>,
## #   hormonal_contraceptives <lgl>, hormonal_contraceptives_years <dbl>,
## #   iud <lgl>, iud_years <dbl>, stds <lgl>, stds_number <dbl>,
## #   stds_condylomatosis <lgl>, stds_cervical_condylomatosis <lgl>,
## #   stds_vaginal_condylomatosis <lgl>,
## #   stds_vulvo_perineal_condylomatosis <lgl>, stds_syphilis <lgl>,
## #   stds_pelvic_inflammatory_disease <lgl>, stds_genital_herpes <lgl>, ...
```

```
summary(cervical)
```

```
##       ...1            age        number_of_sexual_partners
##  Min.   :  1.0   Min.   :13.00   Min.   : 1.000  
##  1st Qu.:215.2   1st Qu.:20.00   1st Qu.: 2.000  
##  Median :429.5   Median :25.00   Median : 2.000  
##  Mean   :429.5   Mean   :26.82   Mean   : 2.512  
##  3rd Qu.:643.8   3rd Qu.:32.00   3rd Qu.: 3.000  
##  Max.   :858.0   Max.   :84.00   Max.   :28.000  
##  first_sexual_intercourse num_of_pregnancies   smokes         smokes_years  
##  Min.   :10.00            Min.   : 0.000     Mode :logical   Min.   : 0.00  
##  1st Qu.:15.00            1st Qu.: 1.000     FALSE:735       1st Qu.: 0.00  
##  Median :17.00            Median : 2.000     TRUE :123       Median : 0.00  
##  Mean   :16.99            Mean   : 2.258                     Mean   : 1.22  
##  3rd Qu.:18.00            3rd Qu.: 3.000                     3rd Qu.: 0.00  
##  Max.   :32.00            Max.   :11.000                     Max.   :37.00  
##  smokes_packs_year hormonal_contraceptives hormonal_contraceptives_years
##  Min.   : 0.0000   Mode :logical           Min.   : 0.000               
##  1st Qu.: 0.0000   FALSE:377               1st Qu.: 0.000               
##  Median : 0.0000   TRUE :481               Median : 1.000               
##  Mean   : 0.4531                           Mean   : 2.224               
##  3rd Qu.: 0.0000                           3rd Qu.: 2.000               
##  Max.   :37.0000                           Max.   :30.000               
##     iud            iud_years          stds          stds_number   
##  Mode :logical   Min.   : 0.0000   Mode :logical   Min.   :0.000  
##  FALSE:775       1st Qu.: 0.0000   FALSE:779       1st Qu.:0.000  
##  TRUE :83        Median : 0.0000   TRUE :79        Median :0.000  
##                  Mean   : 0.4446                   Mean   :0.155  
##                  3rd Qu.: 0.0000                   3rd Qu.:0.000  
##                  Max.   :19.0000                   Max.   :4.000  
##  stds_condylomatosis stds_cervical_condylomatosis stds_vaginal_condylomatosis
##  Mode :logical       Mode :logical                Mode :logical              
##  FALSE:814           FALSE:858                    FALSE:854                  
##  TRUE :44                                         TRUE :4                    
##                                                                              
##                                                                              
##                                                                              
##  stds_vulvo_perineal_condylomatosis stds_syphilis  
##  Mode :logical                      Mode :logical  
##  FALSE:815                          FALSE:840      
##  TRUE :43                           TRUE :18       
##                                                    
##                                                    
##                                                    
##  stds_pelvic_inflammatory_disease stds_genital_herpes
##  Mode :logical                    Mode :logical      
##  FALSE:857                        FALSE:857          
##  TRUE :1                          TRUE :1            
##                                                      
##                                                      
##                                                      
##  stds_molluscum_contagiosum stds_aids         stds_hiv       stds_hepatitis_b
##  Mode :logical              Mode :logical   Mode :logical   Mode :logical   
##  FALSE:857                  FALSE:858       FALSE:840       FALSE:857       
##  TRUE :1                                    TRUE :18        TRUE :1         
##                                                                             
##                                                                             
```

```
##
##    stds_hpv       stds_number_of_diagnosis dx_cancer         dx_cin
##  Mode :logical   Min.   :0.00000          Mode :logical   Mode :logical
##  FALSE:856       1st Qu.:0.00000          FALSE:840       FALSE:849
##  TRUE :2         Median :0.00000          TRUE :18        TRUE :9
##                  Mean   :0.08741
##                  3rd Qu.:0.00000
##                  Max.   :3.00000
##     dx_hpv            dx          hinselmann       schiller
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:840       FALSE:834       FALSE:823       FALSE:784
##  TRUE :18        TRUE :24        TRUE :35        TRUE :74
##
##
##
##    citology         biopsy            age.rz           pregnancy.rz
##  Mode :logical   Mode :logical   Min.   :-3.044808   Min.   :-2.33030
##  FALSE:814       FALSE:803       1st Qu.:-0.721968   1st Qu.:-0.92930
##  TRUE :44        TRUE :55        Median :-0.048167   Median : 0.01459
##                                  Mean   : 0.001286   Mean   : 0.02305
##                                  3rd Qu.: 0.675406   3rd Qu.: 0.70316
##                                  Max.   : 3.044808   Max.   : 3.04481
##  sexualpartners.rz   hormonal.rz
##  Min.   :-1.17255   Min.   :-1.00620
##  1st Qu.:-0.21764   1st Qu.:-1.00620
##  Median :-0.21764   Median : 0.07448
##  Mean   : 0.02593   Mean   : 0.03652
##  3rd Qu.: 0.54866   3rd Qu.: 0.42380
##  Max.   : 3.04481   Max.   : 3.04481
```

After looking over the data, we realized that the data had been read slightly wrong and a column of id numbers was made which is useless in our case. To remove this column, we used the select() function to select all of the columns we wanted which, in this case, was all the columns but the first.

```
cervical <- select(cervical, 2:39)
names(cervical)
```

```
##  [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] "dx_cancer"                          "dx_cin"
## [29] "dx_hpv"                             "dx"
## [31] "hinselmann"                         "schiller"
## [33] "citology"                           "biopsy"
## [35] "age.rz"                             "pregnancy.rz"
## [37] "sexualpartners.rz"                  "hormonal.rz"
```

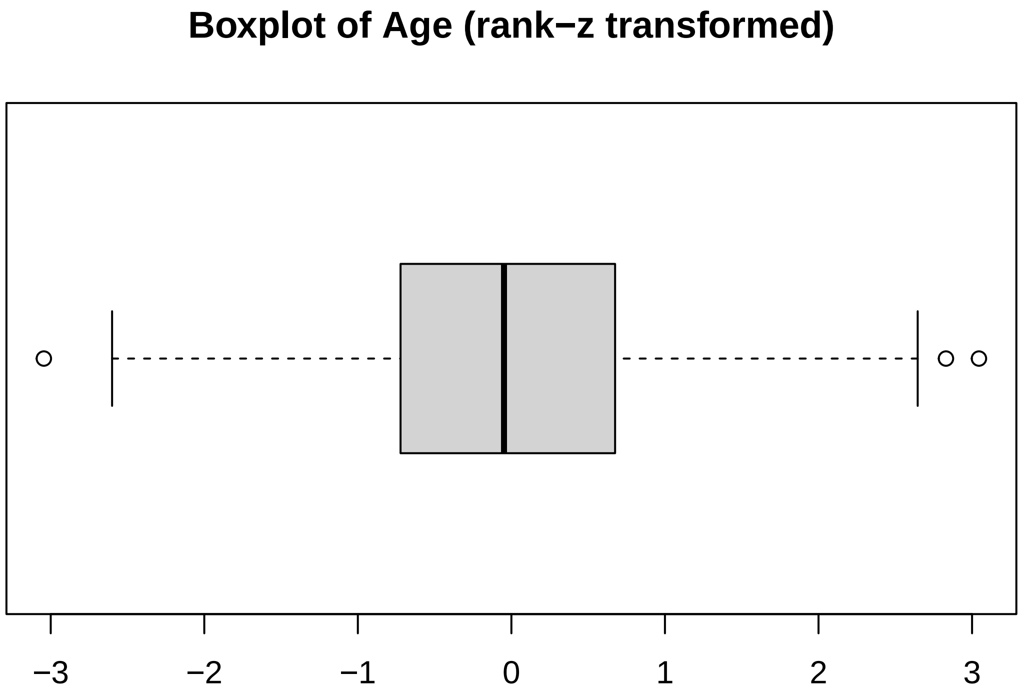
Various Plots and Graphs of Numerical Data

To better help us understand the data, it is important to create various plots and graphs of the variables to most importantly check to see if the data is normally distributed. There are many different sorts of models that we can used to visualize the data, but in this specific case we decided that box plots, histograms, and qq-plots would provide the best descriptions of the data. First we started off by choosing the data that we thought would be the most interesting/give interesting results. For this we chose age (the rank-z transformation), the number of sexual partners (the rank-z transformation), the age at which the patients had their first sexual intercourse, and number of pregnancies (the rank-z transformation). We also decided not to use the par() function since on the final markdown it would make the plots to small to comprehend the data shown on them.

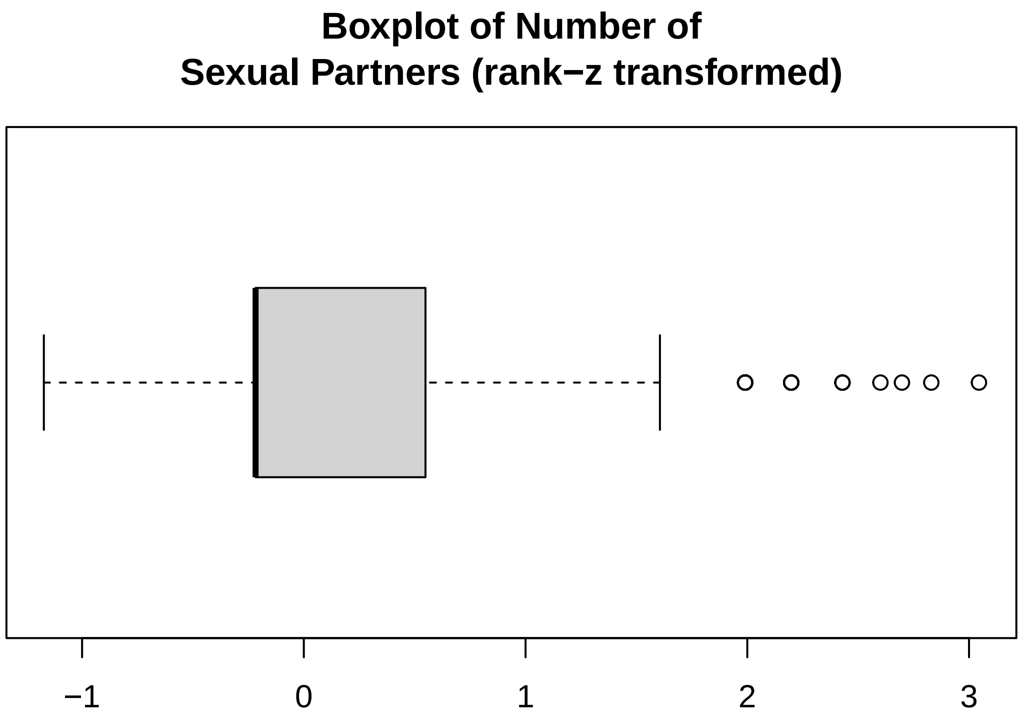
Boxplot

In general, box plots of very helpful in terms of understanding the general spread and range of the data. They also show any outliers in the data. Overall, the spread was good and there weren't many outliers in all the box plots. The one abnormality that we observed was the first quartile and median being the same for Number of Sexual Partners. This was most likely because there were so many occurrences of patients with only one sexual partner. We also included the "horizontal = TRUE" to make the box plots horizontal which makes them more comprehensible.

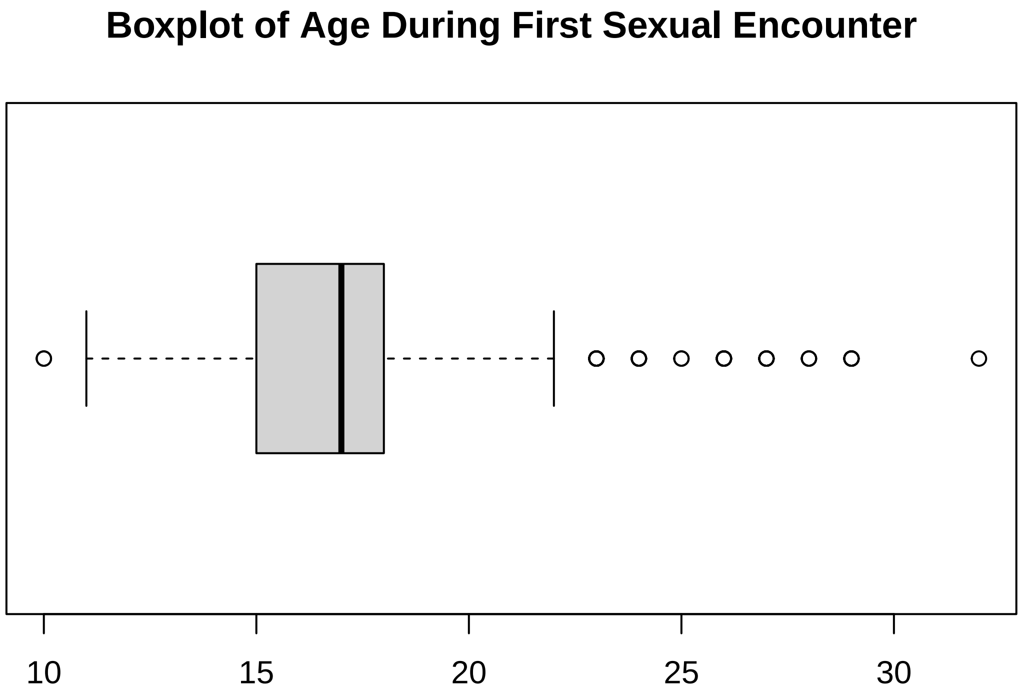
```r
boxplot(cervical$age.rz, horizontal = TRUE, main = "Boxplot of Age (rank-z transformed)")
```



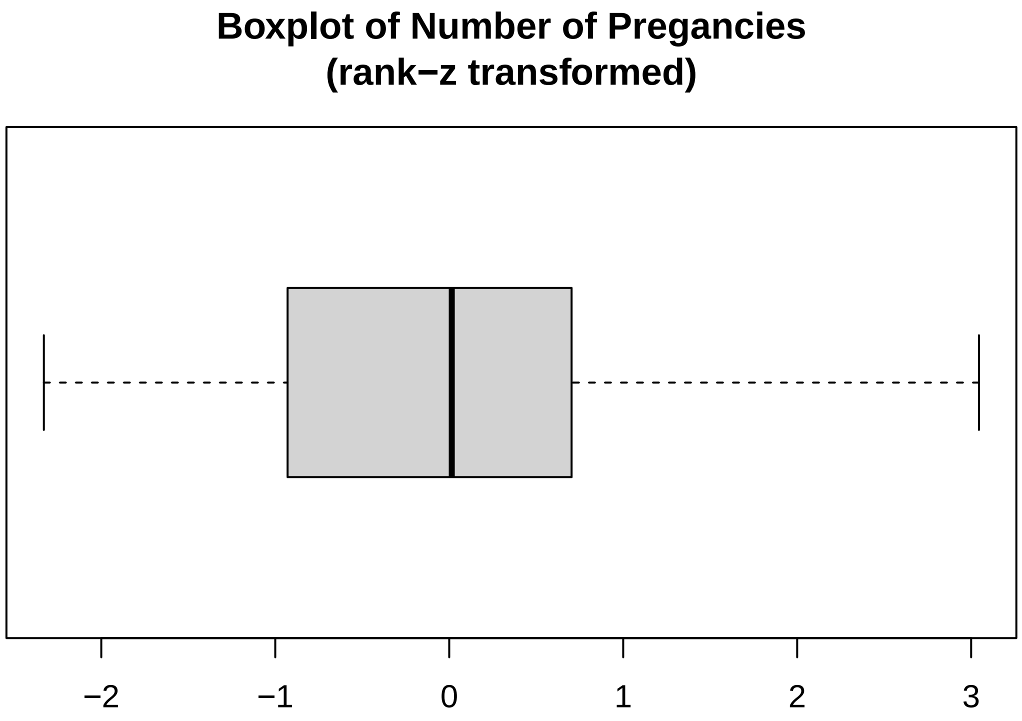
```
boxplot(cervical$sexualpartners.rz, horizontal = TRUE, main = "Boxplot of Number of
Sexual Partners (rank-z transformed)")
```



```
boxplot(cervical$first_sexual_intercourse, horizontal = TRUE, main = "Boxplot of Age During First Sexual
```



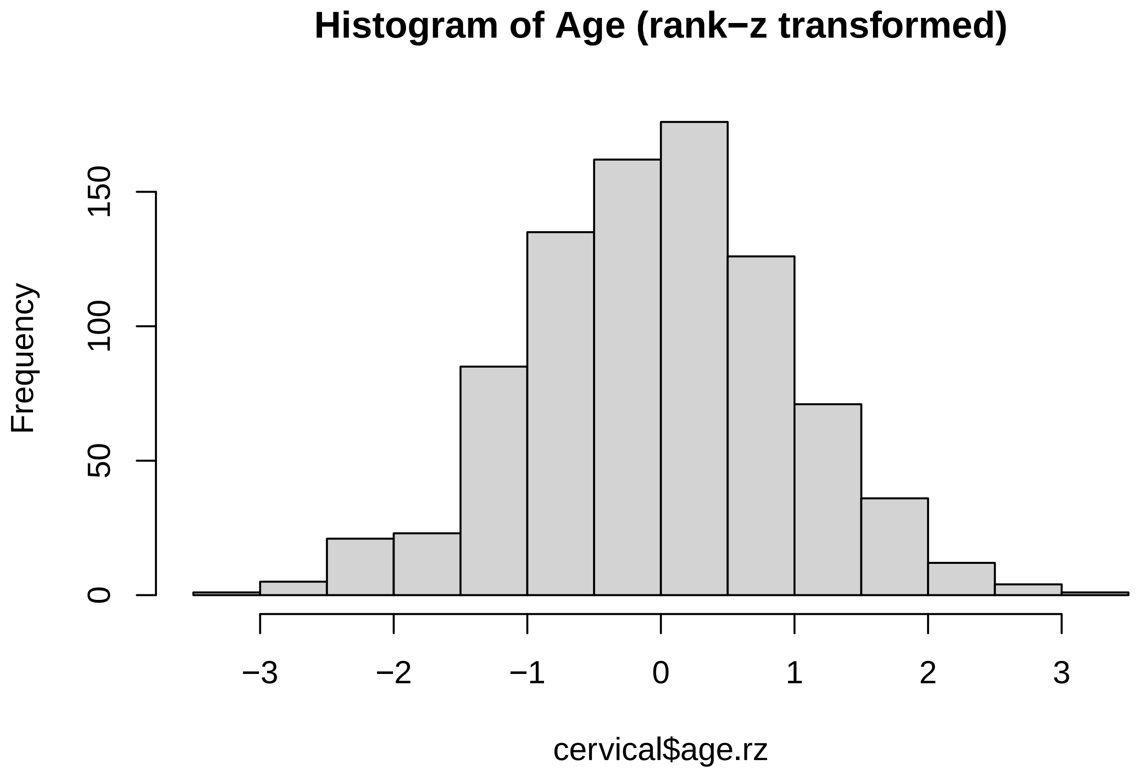
```
boxplot(cervical$pregnancy.rz, horizontal = TRUE, main = "Boxplot of Number of Pregancies
(rank-z transformed)")
```



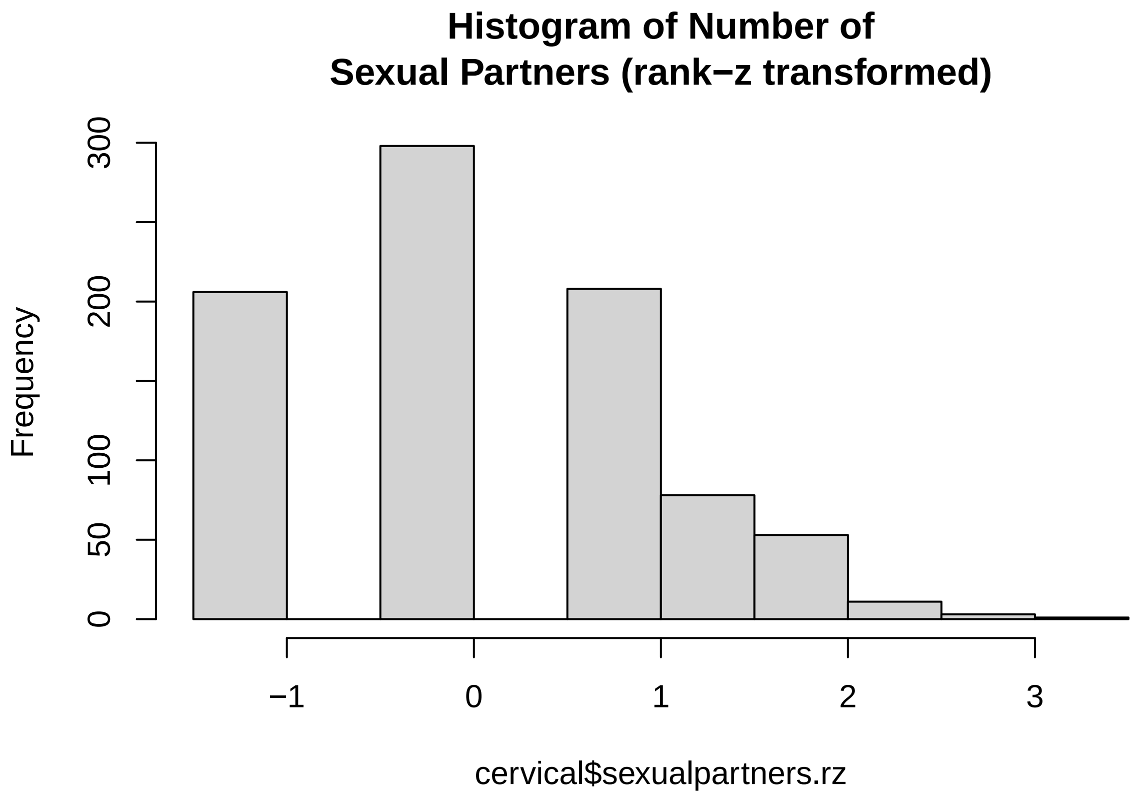
Histogram

Next, we created histogram plots for the same values used in making the box plot. Histograms help us determine the general spread somewhat better than box plots, but don't give us any information on median, range, and IQR. Histograms are also much more helpful in understanding the frequency of each value. To obtain these histograms we used the hist() function and entered in each graph.. All in all, the data seems to be distributed well, especially age, there are some problems especially how sine if the data, mostly the Age During First Sexual Intercourse is skewed right.

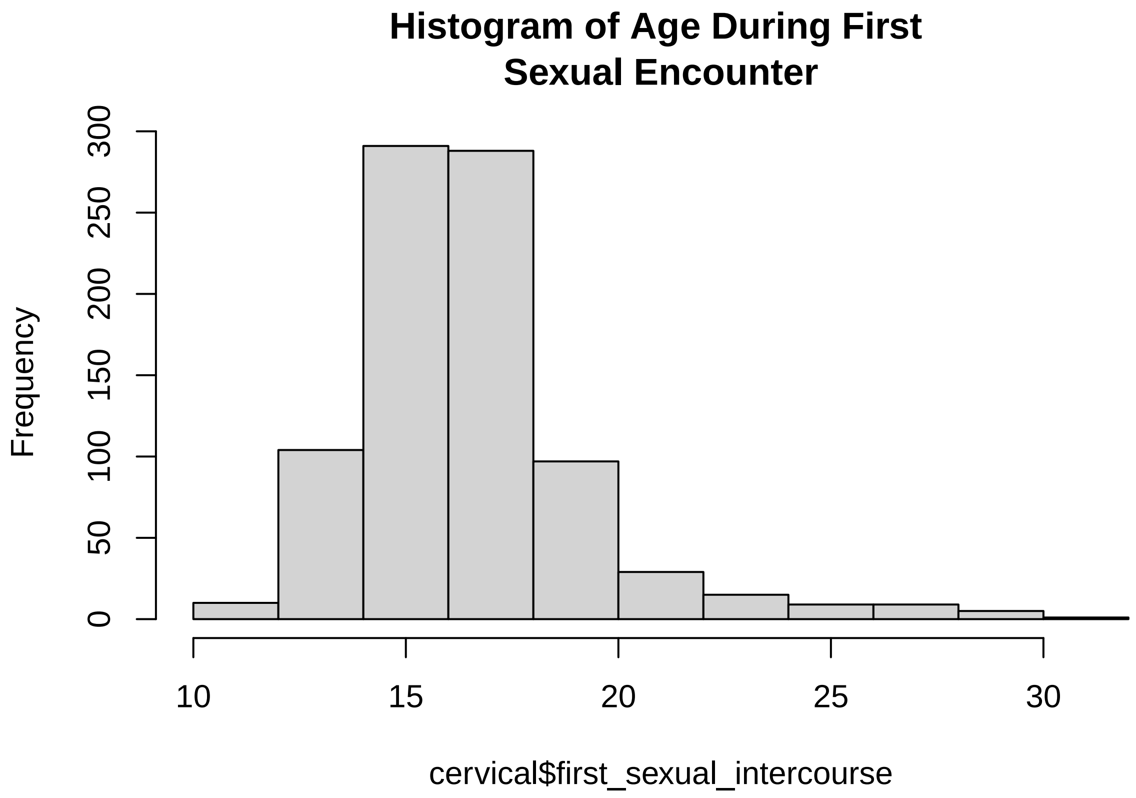
```r
hist(cervical$age.rz, main = "Histogram of Age (rank-z transformed)")
```



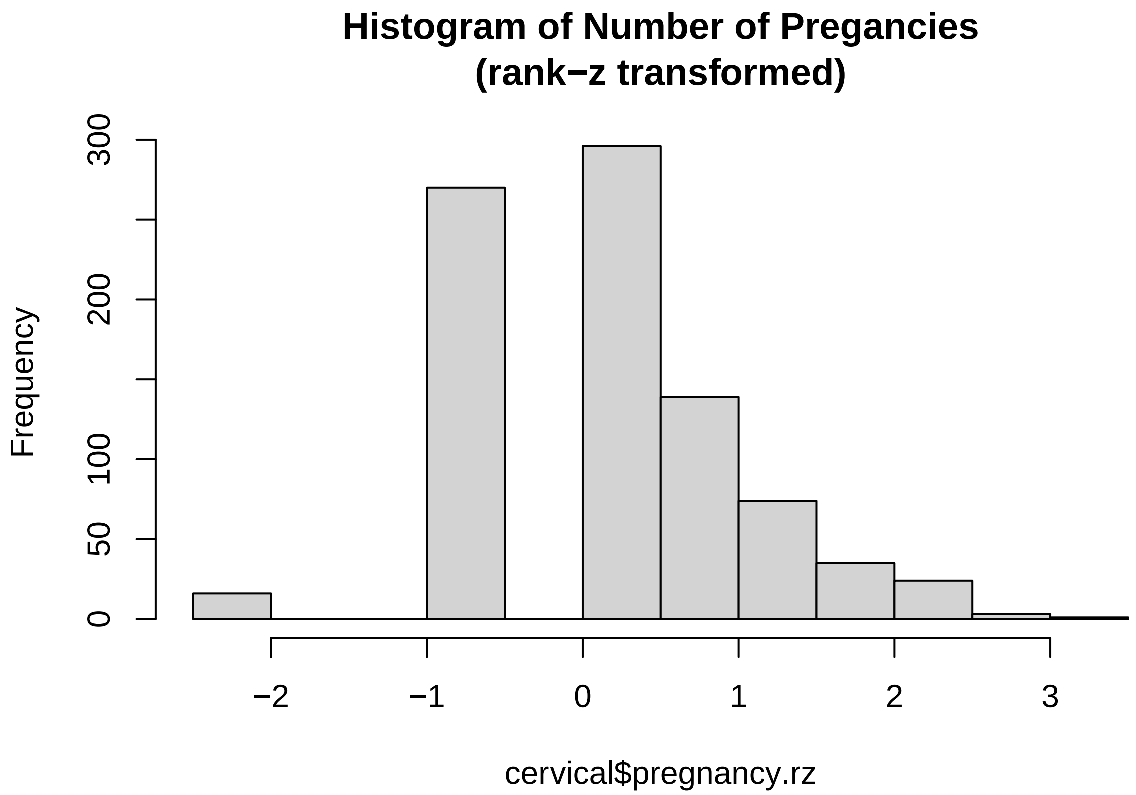
```r
hist(cervical$sexualpartners.rz, main = "Histogram of Number of
Sexual Partners (rank-z transformed)")
```



```
hist(cervical$first_sexual_intercourse, main = "Histogram of Age During First
Sexual Encounter")
```



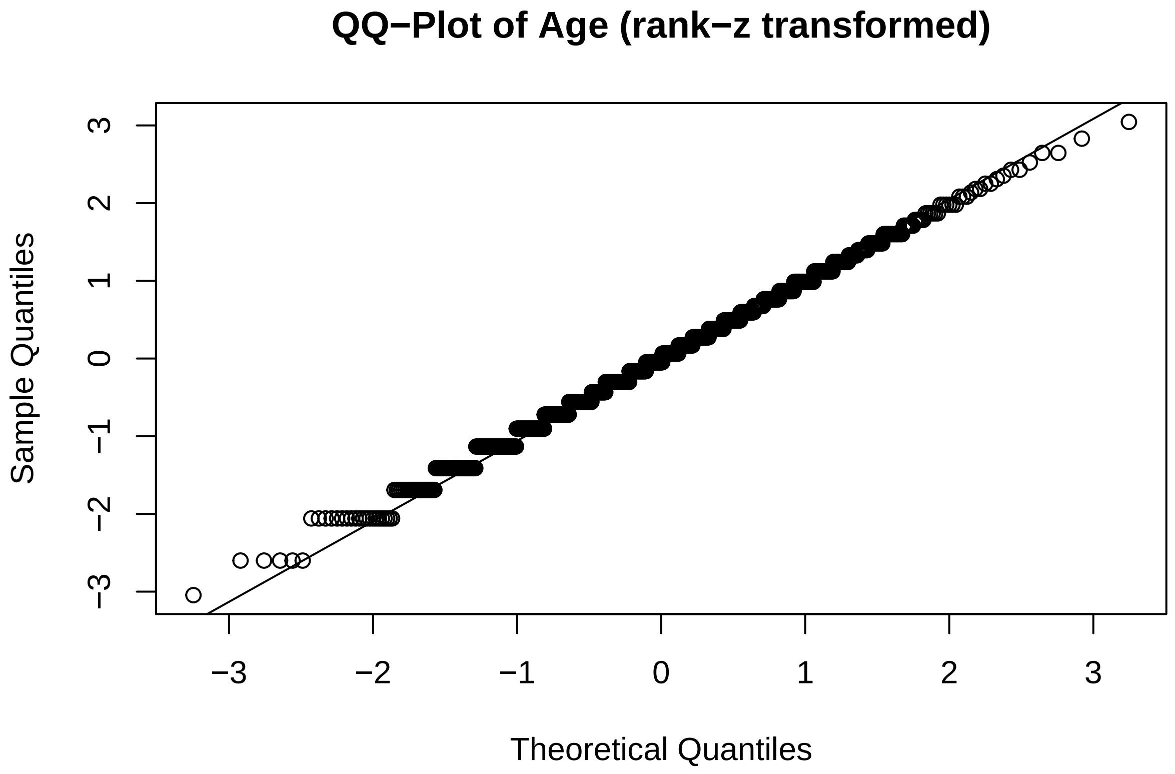
```
hist(cervical$pregnancy.rz, main = "Histogram of Number of Pregancies
(rank-z transformed)")
```



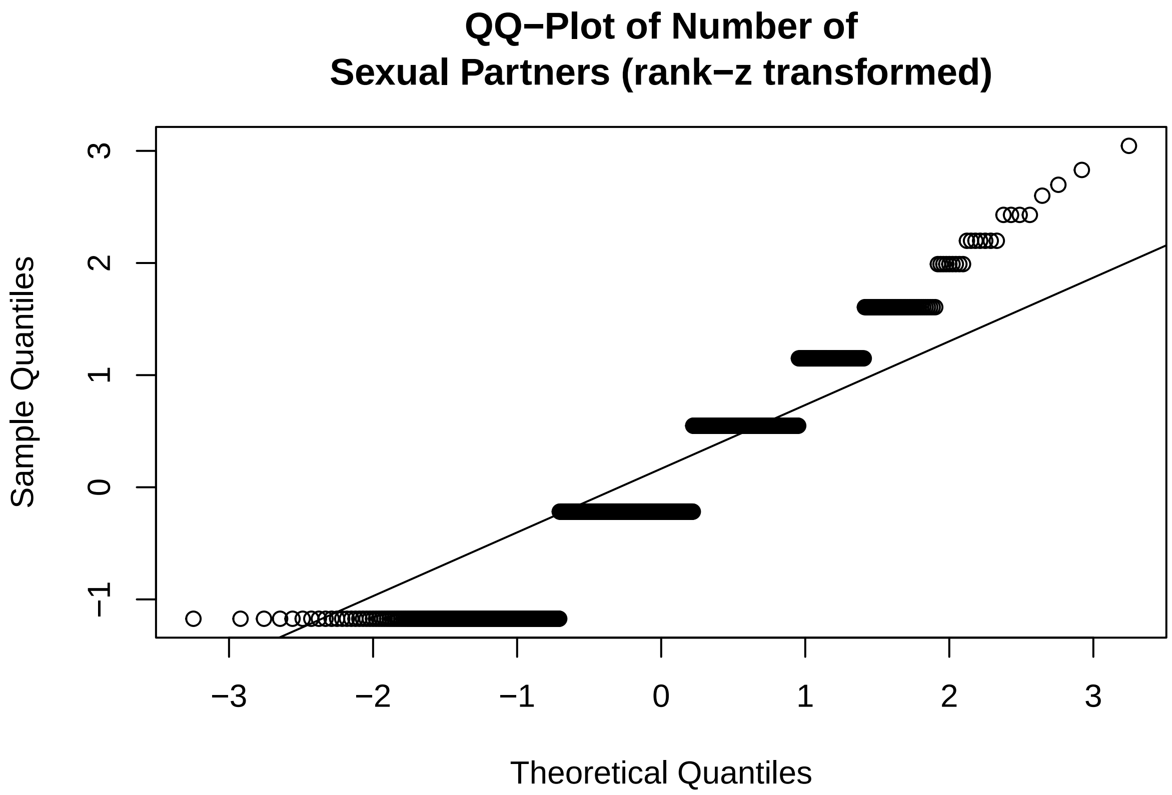
QQ-Plots and QQ-line

Lastly, we used qq-plots, which are models of the theoretical quantities and sample quantities. In general, these are also very helpful in determining which of the variables are normally distributed. To tell if the data is normally distributed, all you must do is look to see how linear the qq-plot is and how well it lines up with the qq-line which is basically a line of best fit for the qq-plot. Overall, the data seems to be normally distributed. The "Number of Sexual Partners" qq-plot seems the least distributed of the group. Additionally, the other data seems to have a lot of values that many people have in common with each other. This makes sense since most people will have similar amounts of sexual partners and very similar amounts of pregnancies, though, the pregnancy data did seem to vary more.

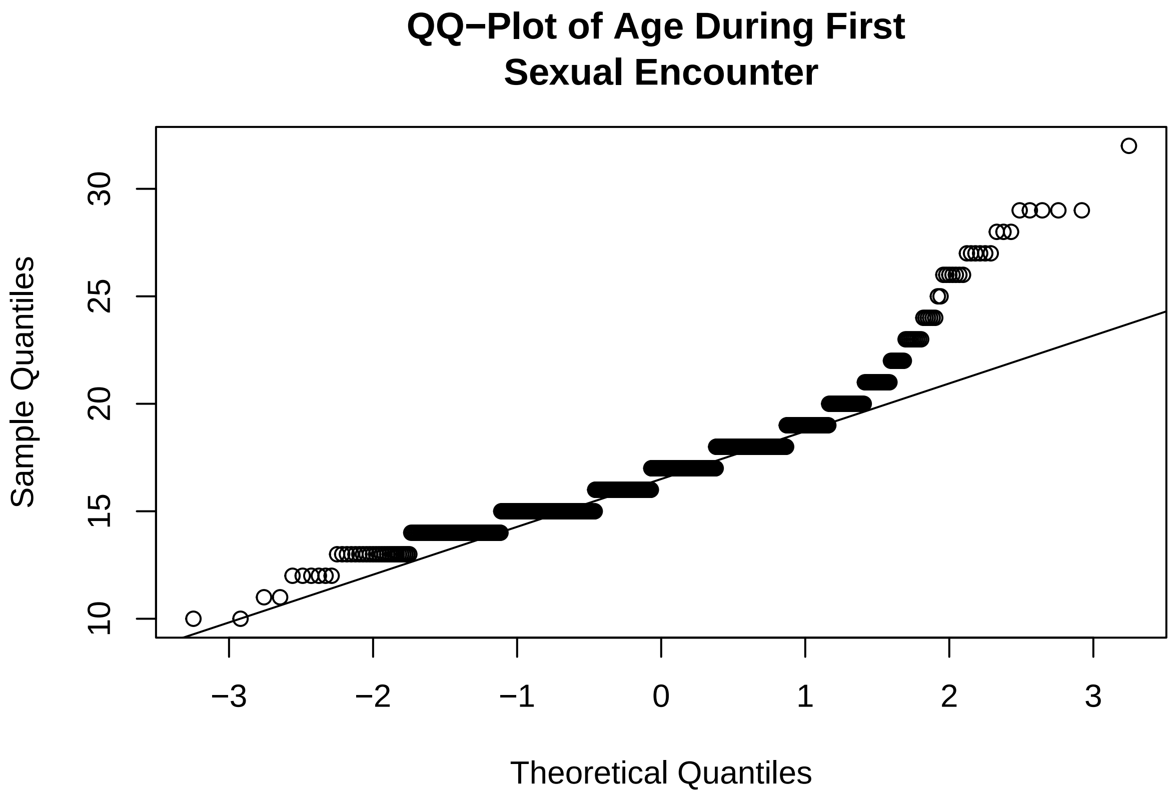
```
qqnorm(cervical$age.rz, main = "QQ-Plot of Age (rank-z transformed)")
qqline(cervical$age.rz)
```



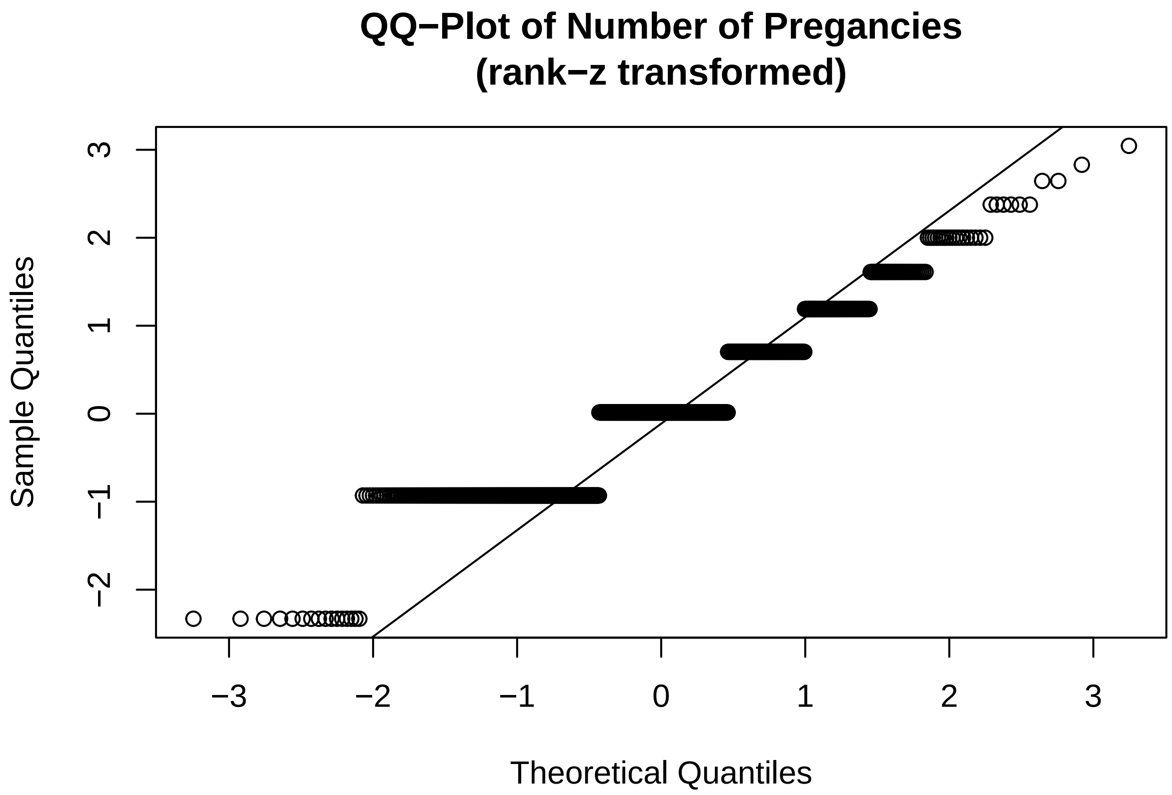
```
qqnorm(cervical$sexualpartners.rz, main = "QQ-Plot of Number of
Sexual Partners (rank-z transformed)")
qqline(cervical$sexualpartners.rz)
```



```
qqnorm(cervical$first_sexual_intercourse, main = "QQ-Plot of Age During First
Sexual Encounter")
qqline(cervical$first_sexual_intercourse)
```



```
qqnorm(cervical$pregnancy.rz, main = "QQ-Plot of Number of Pregancies
(rank-z transformed)")
qqline(cervical$pregnancy.rz)
```



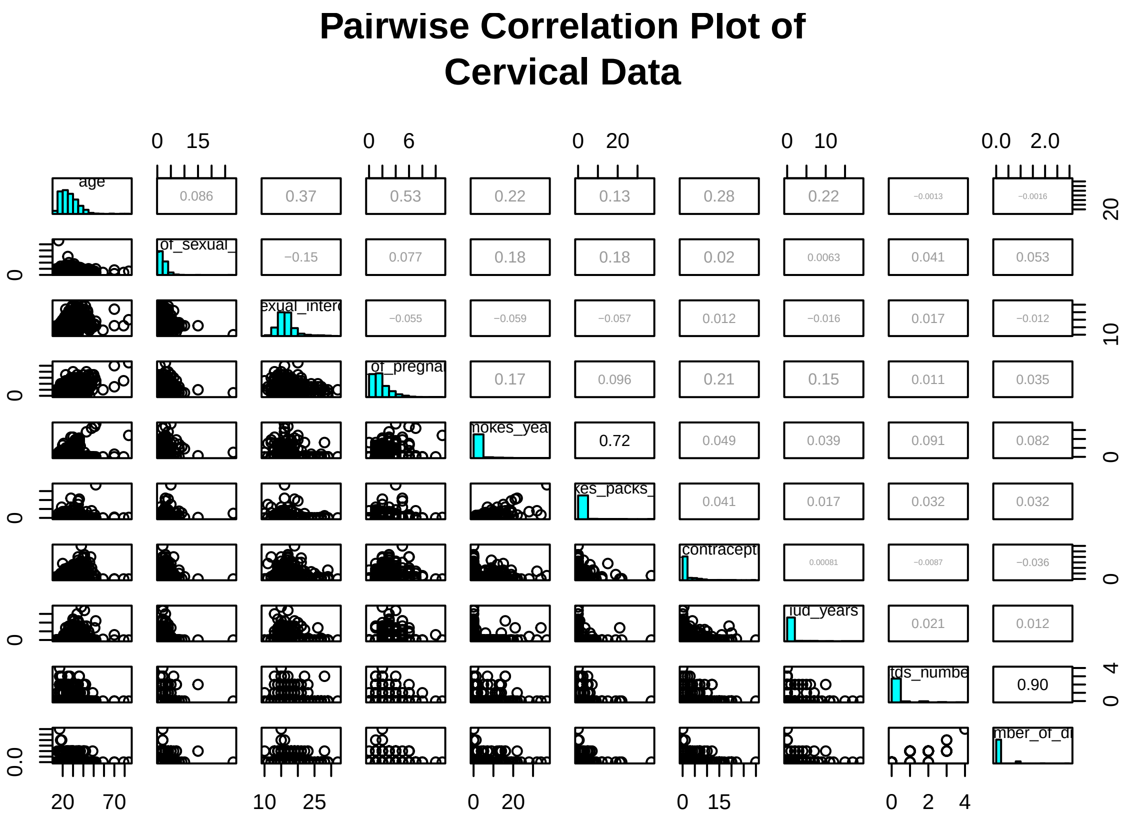
Pairwise Correlation Plot

Once we had created many different models for our data set, we decided that it would be very helpful if we created a pairwise correlation plot for all the numerical data. This pairwise correlation plot is very helpful in understanding the correlation between many of the variables. To create this pairwise correlation plot, we started by using the select() function to create a new data set that only included the numerical values found in the data. Then, we used the pairs() function to create the plot. This correlation plot included scatter plots comparing them, histograms of each, and the correlation coefficient.

```r
num <- select(cervical, c(1:4, 6, 7, 9, 11, 13, 26))
glimpse(num)
```

```
## Rows: 858
## Columns: 10
## $ age                          <dbl> 18, 15, 34, 52, 46, 42, 51, 26, 45, 44, ~
## $ number_of_sexual_partners    <dbl> 4, 1, 1, 5, 3, 3, 3, 1, 1, 3, 3, 1, 4, 2~
## $ first_sexual_intercourse     <dbl> 15, 14, 16, 16, 21, 23, 17, 26, 20, 15, ~
## $ num_of_pregnancies           <dbl> 1, 1, 1, 4, 4, 2, 6, 3, 5, 2, 4, 3, 6, 2~
## $ smokes_years                 <dbl> 0.000000, 0.000000, 0.000000, 37.000000,~
## $ smokes_packs_year            <dbl> 0.0, 0.0, 0.0, 37.0, 0.0, 0.0, 3.4, 0.0,~
## $ hormonal_contraceptives_years <dbl> 0.00, 0.00, 0.00, 3.00, 15.00, 0.00, 0.0~
## $ iud_years                    <dbl> 0, 0, 0, 0, 0, 0, 7, 7, 0, 0, 0, 0, 5, 0~
## $ stds_number                  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ stds_number_of_diagnosis     <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
```

```
pairs(num, upper.panel = panel.cor, diag.panel = panel.hist, main = "Pairwise Correlation Plot of
Cervical Data")
```



Multiple Regression

Multiple regressions are key factors in finding the correlation between one “scalar” variable and multiple “explanatory” variables. This helps us in finding any specific correlations between some of the variables. In this case we created a multiple regression with whether or not they had cancer as the scalar variable and age, number of pregnancy, and the age during their first sexual encounter. In this case there weren’t any correlations observed. To create a multiple regression you use the lm() function and put your scalar variable as the y and the explanatory variables all as the x with plus signs separating them. We used the summary function to provide a summary of the regression.

```
agepregfirstcancer <- lm(cervical$dx_cancer ~ cervical$age.rz + cervical$pregnancy.rz +
cervical$first_sexual_intercourse)
agepregfirstcancer
```

```
##
## Call:
## lm(formula = cervical$dx_cancer ~ cervical$age.rz + cervical$pregnancy.rz +
##     cervical$first_sexual_intercourse)
##
## Coefficients:
##                        (Intercept)                      cervical$age.rz
##                      0.001203                                 0.015835
##             cervical$pregnancy.rz  cervical$first_sexual_intercourse
##                     -0.001384                                 0.001165
```

```
summary(agepregfirstcancer)
```

```
##
## Call:
## lm(formula = cervical$dx_cancer ~ cervical$age.rz + cervical$pregnancy.rz +
##     cervical$first_sexual_intercourse)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.07358 -0.03149 -0.02022 -0.00856  0.99084
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        0.001203   0.034460   0.035   0.9722
## cervical$age.rz                    0.015835   0.006466   2.449   0.0145 *
## cervical$pregnancy.rz             -0.001384   0.006374  -0.217   0.8282
## cervical$first_sexual_intercourse  0.001165   0.002006   0.581   0.5616
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1427 on 854 degrees of freedom
## Multiple R-squared:  0.01364,    Adjusted R-squared:  0.01017
## F-statistic: 3.937 on 3 and 854 DF,  p-value: 0.008346
```

New Data Set Focused On Relationship Between Cervical Cancer, HPV, and Smoking

After we created a multiple linear regression I was still interested to see if there was any correlation between the data since the regression didn't provide much evidence that there was. We decided it would be interesting to see if there were any correlation between cervical cancer, HPV, and smoking. We decided to create a new data set based off the first. To do so we ran the cleaned cervical cancer data through many pipes. First we selected the columns we needed with the select() function. Then we arranged the data from biggest smoker to smallest with the arrange() function. Lastly, we filtered out any data that wasn't a patient with cervical cancer or a patient with HPV. We did this because it helps to show the number of people that had both HPV and Cancer and the number that only had one of the other. This helps in seeing if there were a lot more cancer patients than HPV or vice versa. In the end, the biggest smokers in the data set both had cervical cancer which shows some correlation between both and there were also about 89% of the cancer patients had HPV which matches up with the data given in the article. There were also not many HPV patients without HPV which was surprising.

```
newcerv <- cervical %>% select(age.rz, pregnancy.rz, sexualpartners.rz, smokes, smokes_years, dx_cancer
head(newcerv)
```

```
## # A tibble: 6 x 7
##   age.rz pregnancy.rz sexualpartners.rz smokes smokes_years dx_cancer dx_hpv
##    <dbl>        <dbl>             <dbl> <lgl>         <dbl> <lgl>     <lgl>
## 1  2.43         1.19              1.61  TRUE          37    TRUE      TRUE
## 2  0.675       -0.929             0.549 TRUE          11    TRUE      TRUE
## 3 -0.559       0.703              1.61  TRUE           1.27 FALSE     TRUE
## 4  1.33        0.0146             0.549 FALSE          1.22 TRUE      TRUE
## 5  1.98        1.61              -1.17  FALSE          0    TRUE      TRUE
## 6  1.48        0.0146            -1.17  FALSE          0    TRUE      TRUE
```

Group Summary of Data

There is a very useful tidyverse function called "group_by" which lets us group datasets by a certain variable the summarize it based on that. We decided that it would be interesting to see if there were any trends visible when grouping the data. Created two new data sets both grouped by "dx_cancer". One of the datasets was then summarized by age and age during first sexual encounter and the other was summarized by number of years smoked and the number of packs smoked a year. Overall, we found that people who had cancer were generally older and had sex for the first time when they were older, and also tended to smoke more and smoke more frequently.

```
group_smoke <- cervical %>% group_by(dx_cancer) %>% summarize(meansmoke = mean(smokes_years), meanyears
group_smoke
```

```
## # A tibble: 2 x 3
##   dx_cancer meansmoke meanyearssmoke
##   <lgl>         <dbl>          <dbl>
## 1 FALSE          1.19          0.418
## 2 TRUE           2.73          2.09
```

```
#
group_age <- cervical %>% group_by(dx_cancer) %>% summarize(meanage = mean(age), meanyearagesexual = mea
group_age
```

```
## # A tibble: 2 x 3
##   dx_cancer meanage meanyearagesexual
##   <lgl>       <dbl>             <dbl>
## 1 FALSE        26.7              17.0
## 2 TRUE         33.2              18.3
```

Bayesian Information Criterion

We have taken many looks at the correlation of the data, but we have not covered much of the causation between different variables. In this case, we decided to see which, if any, of the variables had any causality with cervical cancer. We did this by using Bayesian Information Criterion (BIC) models. We first find the causality between one and cervical cancer, then use that number in relation to the other numbers received when calculated causality between smokes, hpv, age, etc. with cervical cancer. If the number obtained when finding the causality between the two variables was 10 or more smaller than that of the number obtained when finding causality between 1 and a variable, then it is deemed that there is potential causality. When calculating to find whether or not there was causality the only variable that we found to cause cervical cancer was HPV.

```
BIC(lm(cervical$dx_cancer~1))
```

```
## [1] -885.2951
```

```
BIC(lm(cervical$dx_cancer~cervical$smokes)) #no causation
```

```
## [1] -878.6962
```

```
BIC(lm(cervical$dx_cancer~cervical$dx_hpv)) #yes lots of causation, but we already knew that
```

```
## [1] -2200.972
```

```
BIC(lm(cervical$dx_cancer~cervical$first_sexual_intercourse)) #no causation
```

```
## [1] -882.4789
```

```
BIC(lm(cervical$dx_cancer~cervical$dx_cin)) #no causation
```

```
## [1] -878.7354
```

```
BIC(lm(cervical$dx_cancer~cervical$age)) #no causation
```

```
## [1] -889.0506
```

```
BIC(lm(cervical$dx_cancer~cervical$number_of_sexual_partners)) #no causation
```

```
## [1] -879.0225
```

```
BIC(lm(cervical$dx_cancer~cervical$num_of_pregnancies)) #no causation
```

```
## [1] -879.7135
```

```
BIC(lm(cervical$dx_cancer~cervical$stds)) #no causation
```

```
## [1] -878.6202
```

Conclusion

This report shows the general steps taken when performing EDA on a dataset. We started by looking over the structure and summary, then we created some models, after that we created some regressions and also dove deeper into other correlation that seemed interesting. Lastly, we ended with finding causality where we used BIC models to calculate this. Overall, in this report, we found that there was some correlation between smoking and having cervical cancer. This is most likely since smoking for a long time and doing so frequently can cause ones immune system to suffer which creates an easier environment for the cancer to develop in. This is also why there was no causality because having an immune system that is weak does not mean that you will have cervical cancer. When using the BIC models we were able to find that HPV does cause cervical cancer, which lines up very well with the article this is based on.