Cervical Cancer Analysis

Srikant Vasudevan

10/1/2019

Abstract

Cervical cancer is a type of cancer that is seen in the cervix, a female reproductive organ. This cancer arises from the malignant and irregular cell growth that can spread to other areas of the body. Symptoms of cervical cancer can start out as not present or negligible, and can lead to bleeding in the vagina as well as sharp pelvic pain.

HPV (human papilliomavirus infection) causes over 90% of cervical cancer cases but there are plenty of other causes that one could attribute to cervical cancer, such as a variety of STDs and STIs. Certain seemingly unrelated things such as smoking, genetics, sexual activity, and the number of pregnancies can also lead to cervical cancer.

Below is a chart consisting of some of the most prominent in the datase and the number of people who have that std every year.

STD Name	# of Cases a Year
Condylomatosis	N/A
Cervical Condylomatosis	N/A
Syphilis	30,600
Pelvic Inflammatory Disease	N/A
Genital Herpes	24,000,000
Molluscum Contagiosum	
AIDS	1,850,000
HIV	1,800,000
Hep. B	20,900
HPV	5,500,000

```
setwd("C:/Users/Srikant/Desktop/Data Science/Week 5/Case Study 2")
source("./myfunctionsaug.R")
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(stringr)
cervical <- read.csv("./ccdataMod.csv")
```

Description of the dataset

To initially analyze the data, we need to look at the descriptive statistics, which will show us things such as if the person has a certain STD of STI,

```
summary(cervical)
```

```
##        X         smokes_years_   smokes_packs_year_      age
##  Min.   :  1.0   Min.   : 0.00   Min.   : 0.0000    Min.   :13.00
##  1st Qu.:215.2   1st Qu.: 0.00   1st Qu.: 0.0000    1st Qu.:20.00
##  Median :429.5   Median : 0.00   Median : 0.0000    Median :25.00
##  Mean   :429.5   Mean   : 1.22   Mean   : 0.4531    Mean   :26.82
##  3rd Qu.:643.8   3rd Qu.: 0.00   3rd Qu.: 0.0000    3rd Qu.:32.00
##  Max.   :858.0   Max.   :37.00   Max.   :37.0000    Max.   :84.00
##  number_of_sexual_partners first_sexual_intercourse num_of_pregnancies
##  Min.   : 1.000            Min.   :10               Min.   : 0.000
##  1st Qu.: 2.000            1st Qu.:15               1st Qu.: 1.000
##  Median : 2.000            Median :17               Median : 2.000
##  Mean   : 2.528            Mean   :17               Mean   : 2.276
##  3rd Qu.: 3.000            3rd Qu.:18               3rd Qu.: 3.000
##  Max.   :28.000            Max.   :32               Max.   :11.000
##  hormonal_contraceptives_years_   iud_years_       stds_number_
##  Min.   : 0.000                 Min.   : 0.0000   Min.   :0.0000
##  1st Qu.: 0.000                 1st Qu.: 0.0000   1st Qu.:0.0000
##  Median : 1.000                 Median : 0.0000   Median :0.0000
##  Mean   : 2.256                 Mean   : 0.5148   Mean   :0.1766
##  3rd Qu.: 2.256                 3rd Qu.: 0.0000   3rd Qu.:0.0000
##  Max.   :30.000                 Max.   :19.0000   Max.   :4.0000
##  stds_number_of_diagnosis stds_time_since_first_diagnosis
##  Min.   :0.00000          Min.   : 1.000
##  1st Qu.:0.00000          1st Qu.: 6.141
##  Median :0.00000          Median : 6.141
```

```
##  Mean   :0.08741          Mean   : 6.141
##  3rd Qu.:0.00000          3rd Qu.: 6.141
##  Max.   :3.00000          Max.   :22.000
##  stds_time_since_last_diagnosis   smokes        hormonal_contraceptives
##  Min.   : 1.000                 Mode :logical   Mode :logical
##  1st Qu.: 5.817                 FALSE:722       FALSE:269
##  Median : 5.817                 TRUE :136       TRUE :589
##  Mean   : 5.817
##  3rd Qu.: 5.817
##  Max.   :22.000
##     iud             stds         stds_condylomatosis
##  Mode :logical   Mode :logical   Mode :logical
##  FALSE:658       FALSE:674       FALSE:709
##  TRUE :200       TRUE :184       TRUE :149
##
##
##
##  stds_cervical_condylomatosis stds_vaginal_condylomatosis
##  Mode :logical                Mode :logical
##  FALSE:753                    FALSE:749
##  TRUE :105                    TRUE :109
##
##
##
##  stds_vulvo_perineal_condylomatosis stds_syphilis
##  Mode :logical                      Mode :logical
##  FALSE:710                          FALSE:735
##  TRUE :148                          TRUE :123
##
##
##
##  stds_pelvic_inflammatory_disease stds_genital_herpes
##  Mode :logical                    Mode :logical
##  FALSE:752                        FALSE:752
##  TRUE :106                        TRUE :106
##
##
##
##  stds_molluscum_contagiosum stds_aids        stds_hiv
##  Mode :logical              Mode :logical   Mode :logical
##  FALSE:752                  FALSE:753       FALSE:735
##  TRUE :106                  TRUE :105       TRUE :123
##
##
##
##  stds_hepatitis_b  stds_hpv       dx_cancer         dx_cin
##  Mode :logical    Mode :logical   Mode :logical   Mode :logical
##  FALSE:752        FALSE:751       FALSE:840       FALSE:849
##  TRUE :106        TRUE :107       TRUE :18        TRUE :9
##
##
##
##    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
##  Mode :logical   Mode :logical
##  FALSE:814       FALSE:803
##  TRUE :44        TRUE :55
##
##
##
```

```
dim(cervical)
```

```
## [1] 858  37
```

```
str(cervical)
```

```
## 'data.frame':    858 obs. of  37 variables:
##  $ X                                 : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ smokes_years_                     : num  0 0 0 37 0 ...
##  $ smokes_packs_year_                : num  0 0 0 37 0 0 3.4 0 0 2.8 ...
##  $ age                               : int  18 15 34 52 46 42 51 26 45 44 ...
##  $ number_of_sexual_partners         : num  4 1 1 5 3 3 3 1 1 3 ...
##  $ first_sexual_intercourse          : num  15 14 17 16 21 ...
##  $ num_of_pregnancies                : num  1 1 1 4 4 ...
##  $ hormonal_contraceptives_years_    : num  0 0 0 3 15 0 0 2 0 0 ...
##  $ iud_years_                        : num  0 0 0 0 0 ...
##  $ stds_number_                      : num  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   : num  6.14 6.14 6.14 6.14 6.14 ...
##  $ stds_time_since_last_diagnosis    : num  5.82 5.82 5.82 5.82 5.82 ...
##  $ smokes                            : logi  FALSE FALSE FALSE TRUE FALSE FALSE ...
##  $ hormonal_contraceptives           : logi  FALSE FALSE FALSE TRUE TRUE FALSE ...
##  $ iud                               : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ stds                              : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ stds_condylomatosis               : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ stds_cervical_condylomatosis      : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ stds_vaginal_condylomatosis       : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ stds_vulvo_perineal_condylomatosis: logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ stds_syphilis                     : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ stds_pelvic_inflammatory_disease  : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ stds_genital_herpes               : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ stds_molluscum_contagiosum        : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ stds_aids                         : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ stds_hiv                          : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ stds_hepatitis_b                  : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ stds_hpv                          : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ 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 ...
##  $ citology                          : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ biopsy                            : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
```

```
glimpse(cervical)
```

```
## Observations: 858
## Variables: 37
## $ X                                 <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, ...
## $ smokes_years_                     <dbl> 0.000000, 0.000000, 0.00000...
## $ smokes_packs_year_                <dbl> 0.0, 0.0, 0.0, 37.0, 0.0, 0...
## $ age                               <int> 18, 15, 34, 52, 46, 42, 51,...
## $ number_of_sexual_partners         <dbl> 4, 1, 1, 5, 3, 3, 3, 1, 1, ...
## $ first_sexual_intercourse          <dbl> 15.0000, 14.0000, 16.9953, ...
## $ num_of_pregnancies                <dbl> 1.000000, 1.000000, 1.00000...
## $ hormonal_contraceptives_years_    <dbl> 0.00, 0.00, 0.00, 3.00, 15....
## $ iud_years_                        <dbl> 0.0000000, 0.0000000, 0.000...
## $ stds_number_                      <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ stds_number_of_diagnosis          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ stds_time_since_first_diagnosis   <dbl> 6.140845, 6.140845, 6.14084...
## $ stds_time_since_last_diagnosis    <dbl> 5.816901, 5.816901, 5.81690...
## $ smokes                            <lgl> FALSE, FALSE, FALSE, TRUE, ...
## $ hormonal_contraceptives           <lgl> FALSE, FALSE, FALSE, TRUE, ...
## $ iud                               <lgl> FALSE, FALSE, FALSE, FALSE,...
## $ stds                              <lgl> FALSE, FALSE, FALSE, FALSE,...
## $ stds_condylomatosis               <lgl> FALSE, FALSE, FALSE, FALSE,...
## $ stds_cervical_condylomatosis      <lgl> FALSE, FALSE, FALSE, FALSE,...
## $ stds_vaginal_condylomatosis       <lgl> FALSE, FALSE, FALSE, FALSE,...
## $ stds_vulvo_perineal_condylomatosis <lgl> FALSE, FALSE, FALSE, FALSE,...
## $ stds_syphilis                     <lgl> FALSE, FALSE, FALSE, FALSE,...
## $ stds_pelvic_inflammatory_disease  <lgl> FALSE, FALSE, FALSE, FALSE,...
## $ stds_genital_herpes               <lgl> FALSE, FALSE, FALSE, FALSE,...
## $ stds_molluscum_contagiosum        <lgl> FALSE, FALSE, FALSE, FALSE,...
## $ stds_aids                         <lgl> FALSE, FALSE, FALSE, FALSE,...
## $ stds_hiv                          <lgl> FALSE, FALSE, FALSE, FALSE,...
## $ stds_hepatitis_b                  <lgl> FALSE, FALSE, FALSE, FALSE,...
## $ stds_hpv                          <lgl> FALSE, FALSE, FALSE, FALSE,...
## $ dx_cancer                         <lgl> FALSE, FALSE, FALSE, TRUE, ...
## $ dx_cin                            <lgl> FALSE, FALSE, FALSE, FALSE,...
## $ dx_hpv                            <lgl> FALSE, FALSE, FALSE, TRUE, ...
## $ dx                                <lgl> FALSE, FALSE, FALSE, FALSE,...
## $ hinselmann                        <lgl> FALSE, FALSE, FALSE, FALSE,...
## $ schiller                          <lgl> FALSE, FALSE, FALSE, FALSE,...
## $ citology                          <lgl> FALSE, FALSE, FALSE, FALSE,...
## $ biopsy                            <lgl> FALSE, FALSE, FALSE, FALSE,...
```

```
head(cervical)
```

```
##   X smokes_years_ smokes_packs_year_ age number_of_sexual_partners
## 1 1             0                  0  18                          4
## 2 2             0                  0  15                          1
## 3 3             0                  0  34                          1
```

```
## 4 4            37                37  52                        5
## 5 5             0                 0  46                        3
## 6 6             0                 0  42                        3
##   first_sexual_intercourse num_of_pregnancies
## 1                  15.0000                 1
## 2                  14.0000                 1
## 3                  16.9953                 1
## 4                  16.0000                 4
## 5                  21.0000                 4
## 6                  23.0000                 2
##   hormonal_contraceptives_years_ iud_years_ stds_number_
## 1                              0          0            0
## 2                              0          0            0
## 3                              0          0            0
## 4                              3          0            0
## 5                             15          0            0
## 6                              0          0            0
##   stds_number_of_diagnosis stds_time_since_first_diagnosis
## 1                        0                        6.140845
## 2                        0                        6.140845
## 3                        0                        6.140845
## 4                        0                        6.140845
## 5                        0                        6.140845
## 6                        0                        6.140845
##   stds_time_since_last_diagnosis smokes hormonal_contraceptives   iud
## 1                       5.816901  FALSE                   FALSE FALSE
## 2                       5.816901  FALSE                   FALSE FALSE
## 3                       5.816901  FALSE                   FALSE FALSE
## 4                       5.816901   TRUE                    TRUE FALSE
## 5                       5.816901  FALSE                    TRUE FALSE
## 6                       5.816901  FALSE                   FALSE FALSE
##    stds stds_condylomatosis stds_cervical_condylomatosis
## 1 FALSE              FALSE                        FALSE
## 2 FALSE              FALSE                        FALSE
## 3 FALSE              FALSE                        FALSE
## 4 FALSE              FALSE                        FALSE
## 5 FALSE              FALSE                        FALSE
## 6 FALSE              FALSE                        FALSE
##   stds_vaginal_condylomatosis stds_vulvo_perineal_condylomatosis
## 1                       FALSE                              FALSE
## 2                       FALSE                              FALSE
## 3                       FALSE                              FALSE
## 4                       FALSE                              FALSE
## 5                       FALSE                              FALSE
## 6                       FALSE                              FALSE
##   stds_syphilis stds_pelvic_inflammatory_disease stds_genital_herpes
## 1         FALSE                           FALSE                FALSE
## 2         FALSE                           FALSE                FALSE
## 3         FALSE                           FALSE                FALSE
## 4         FALSE                           FALSE                FALSE
## 5         FALSE                           FALSE                FALSE
## 6         FALSE                           FALSE                FALSE
##   stds_molluscum_contagiosum stds_aids stds_hiv stds_hepatitis_b stds_hpv
## 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
##   dx_cancer dx_cin dx_hpv    dx hinselmann schiller citology 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
```

```
names(cervical)
```

```
##  [1] "X"
##  [2] "smokes_years_"
##  [3] "smokes_packs_year_"
##  [4] "age"
##  [5] "number_of_sexual_partners"
##  [6] "first_sexual_intercourse"
##  [7] "num_of_pregnancies"
##  [8] "hormonal_contraceptives_years_"
##  [9] "iud_years_"
## [10] "stds_number_"
## [11] "stds_number_of_diagnosis"
## [12] "stds_time_since_first_diagnosis"
## [13] "stds_time_since_last_diagnosis"
## [14] "smokes"
## [15] "hormonal_contraceptives"
## [16] "iud"
## [17] "stds"
## [18] "stds_condylomatosis"
## [19] "stds_cervical_condylomatosis"
## [20] "stds_vaginal_condylomatosis"
## [21] "stds_vulvo_perineal_condylomatosis"
## [22] "stds_syphilis"
## [23] "stds_pelvic_inflammatory_disease"
## [24] "stds_genital_herpes"
## [25] "stds_molluscum_contagiosum"
## [26] "stds_aids"
## [27] "stds_hiv"
## [28] "stds_hepatitis_b"
## [29] "stds_hpv"
## [30] "dx_cancer"
## [31] "dx_cin"
## [32] "dx_hpv"
## [33] "dx"
## [34] "hinselmann"
## [35] "schiller"
## [36] "citology"
## [37] "biopsy"
```

The summary statistics tell us that we are dealing with 858 different subjects (rows in the dataset) and 37

variables for each observation (columns in the dataset). The data is primarily logical data (yes-no, true-false), specifically 24 rows.

The X variable measured is able to be ignored as it just numbers the subjects from 1-858. ## Arranging the Dataset

I am going to arrange the dataset by age, this will make it easier for us to do initial analysis on the data

```
cervical <- arrange(cervical, age)
head(cervical)
```

```
##     X smokes_years_ smokes_packs_year_ age number_of_sexual_partners
## 1 673             0                  0  13                  1.000000
## 2 460             0                  0  14                  2.000000
## 3 461             0                  0  14                  2.527644
## 4 464             0                  0  14                  1.000000
## 5 813             0                  0  14                  5.000000
## 6 820             0                  0  14                  1.000000
##   first_sexual_intercourse num_of_pregnancies
## 1                       13           0.000000
## 2                       14           1.000000
## 3                       14           1.000000
## 4                       14           2.000000
## 5                       16           2.275561
## 6                       14           2.275561
##   hormonal_contraceptives_years_ iud_years_ stds_number_
## 1                       0.000000  0.0000000    0.0000000
## 2                       0.000000  0.0000000    0.0000000
## 3                       2.256419  0.5148043    0.1766268
## 4                       0.000000  0.0000000    0.0000000
## 5                       0.080000  0.0000000    0.0000000
## 6                       2.256419  0.0000000    0.0000000
##   stds_number_of_diagnosis stds_time_since_first_diagnosis
## 1                        0                        6.140845
## 2                        0                        6.140845
## 3                        0                        6.140845
## 4                        0                        6.140845
## 5                        0                        6.140845
## 6                        0                        6.140845
##   stds_time_since_last_diagnosis smokes hormonal_contraceptives   iud
## 1                       5.816901  FALSE                   FALSE FALSE
## 2                       5.816901  FALSE                   FALSE FALSE
## 3                       5.816901  FALSE                    TRUE  TRUE
## 4                       5.816901  FALSE                   FALSE FALSE
## 5                       5.816901  FALSE                    TRUE FALSE
## 6                       5.816901  FALSE                    TRUE FALSE
##    stds stds_condylomatosis stds_cervical_condylomatosis
## 1 FALSE              FALSE                        FALSE
## 2 FALSE              FALSE                        FALSE
## 3  TRUE               TRUE                         TRUE
## 4 FALSE              FALSE                        FALSE
## 5 FALSE              FALSE                        FALSE
## 6 FALSE              FALSE                        FALSE
##   stds_vaginal_condylomatosis stds_vulvo_perineal_condylomatosis
## 1                       FALSE                             FALSE
```

```
## 2                       FALSE                               FALSE
## 3                        TRUE                                TRUE
## 4                       FALSE                               FALSE
## 5                       FALSE                               FALSE
## 6                       FALSE                               FALSE
##   stds_syphilis stds_pelvic_inflammatory_disease stds_genital_herpes
## 1         FALSE                           FALSE                FALSE
## 2         FALSE                           FALSE                FALSE
## 3          TRUE                            TRUE                 TRUE
## 4         FALSE                           FALSE                FALSE
## 5         FALSE                           FALSE                FALSE
## 6         FALSE                           FALSE                FALSE
##   stds_molluscum_contagiosum stds_aids stds_hiv stds_hepatitis_b stds_hpv
## 1                      FALSE     FALSE    FALSE            FALSE    FALSE
## 2                      FALSE     FALSE    FALSE            FALSE    FALSE
## 3                       TRUE      TRUE     TRUE             TRUE     TRUE
## 4                      FALSE     FALSE    FALSE            FALSE    FALSE
## 5                      FALSE     FALSE    FALSE            FALSE    FALSE
## 6                      FALSE     FALSE    FALSE            FALSE    FALSE
##   dx_cancer dx_cin dx_hpv    dx hinselmann schiller citology 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     FALSE  FALSE  FALSE FALSE      FALSE    FALSE    FALSE  FALSE
## 5     FALSE  FALSE  FALSE FALSE      FALSE    FALSE    FALSE  FALSE
## 6     FALSE  FALSE  FALSE FALSE      FALSE    FALSE    FALSE  FALSE
```

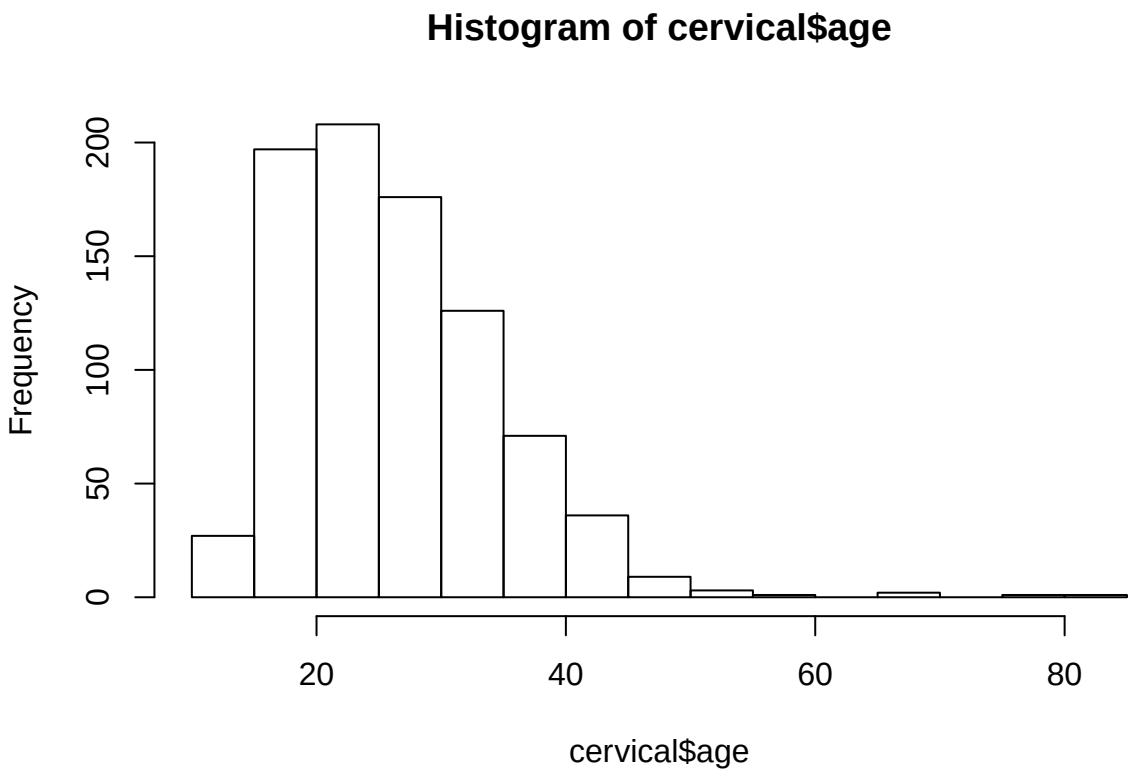
Basic Plotting

Basic plots and graphs in R allow us to visualize the data, to see what shapes and patterns are present with each variable, and ultimately to understand the primary causes and reasons for cervical cancer

Age Data

First let's plot the age, to see what our distribution will look like

```
hist(cervical$age)
```

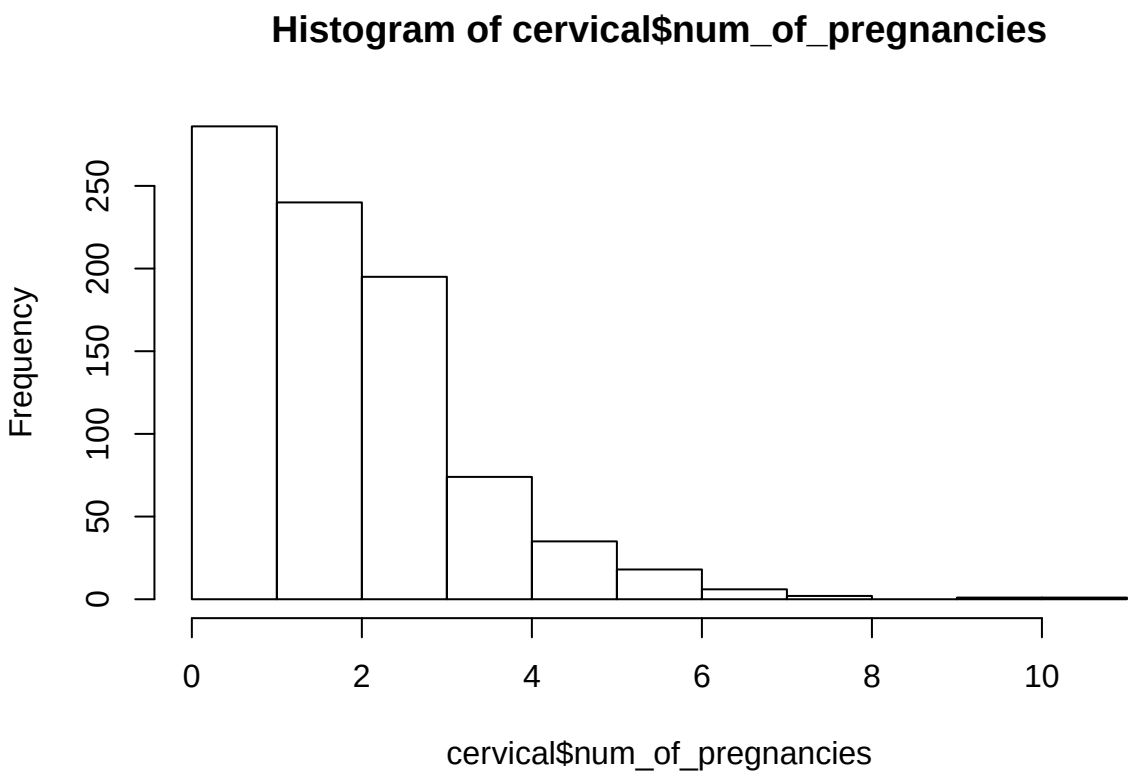


The age data shows that the mean age is somewhere around 25-30 and that the "age distribution" is skewed right, showing that there are less elderly people in this data.

Pregnancy Data

As stated earlier in the abstract, multiple pregnancies allegedly increases the risk of developing cervical cancer. This histogram below will not prove or disprove that statement just yet, but it will instead give us insight on the distribution of the number of pregnancies within the dataset.

```
hist(cervical$num_of_pregnancies)
```

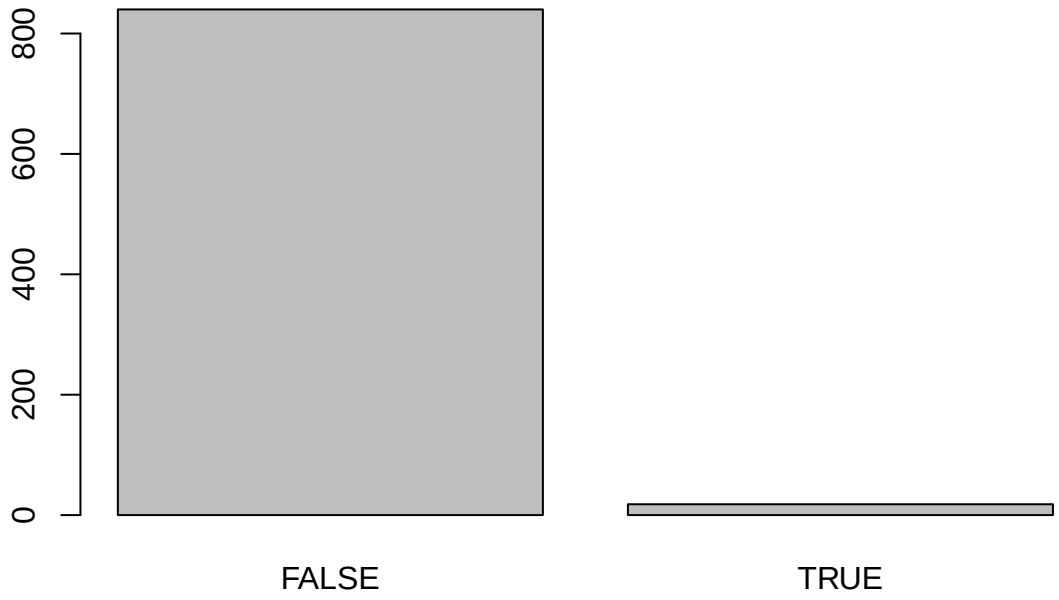


As we can see, similar to the age graph, the distribution is skewed right, with a mean around 2 pregnancies.

Cancer Diagnosis Plot

The last histogram we want to look at, is obviously, the distribution of those who have cervical cancer or don't.

```
barplot(table(cervical$dx_cancer))
```



As clearly seen by the cervical cancer data, there is an almost negligible (but not quite) amount of subjects that had cervical cancer. We will analyze the relationshipt with these data later in the report.

Linear Models and Regression

Linear models and regression allow us to assume certain attributes of those who have cervical cancer, one of the most prominent being causation or increased likelihood in getting cervical cancer.

Age and Number of STDs

Since the cervical cancer data is in logical form, we can't directly correlate it to anything in linear form, but we can look at other relationships which in turn will tell us more about potential causes

```
analysisstdss <- lm(cervical$stds_number_of_diagnosis~cervical$age)
BIC(analysisstdss)
```

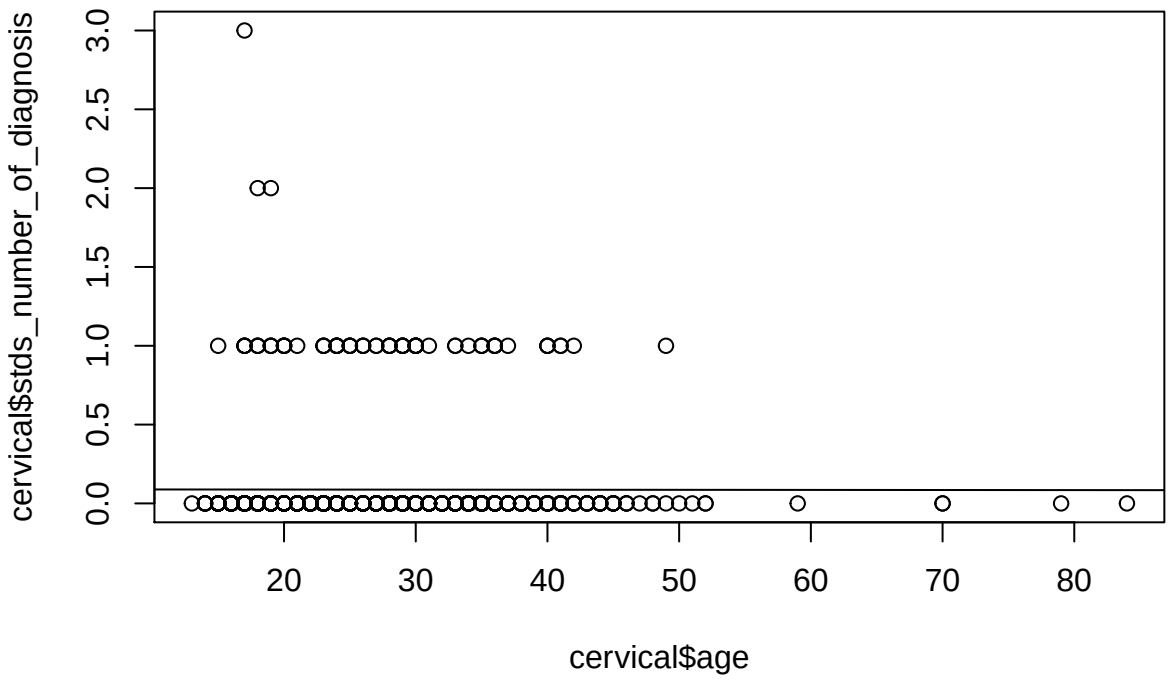
```
## [1] 402.6367
```

```
analysisstdss
```

```
##
## Call:
## lm(formula = cervical$stds_number_of_diagnosis ~ cervical$age)
```

```
##
## Coefficients:
##  (Intercept)  cervical$age
##    8.895e-02    -5.717e-05
```

```
plot(cervical$stds_number_of_diagnosis~cervical$age)
abline(analysisstdss)
```



According to the data and the plot there is little to no correlation or linear regression of the number of std diagnoses in terms of age. This means that more or less of age has little to no effect on the number of STD diagnoses in women.

Cervical Cancer

The causes of cervical cancer have been touched on and analyzed, but the meat of the data, the diagnoses of cervical cancer itself, were not touched on that much. In this for loop, I will be calculating the means of the number of pregnancies in those who did NOT have cervical cancer and those who DO have cervical cancer to determine some sort of statistical significance.

```
value <- 0
number <- 0
value2 <- 0
number2 <- 0
#Column 30 is the column indicating whether the subject has
#cancer and column 7 is the column indicating the number of pregnancies of the subject
for(i in 1:858){
```

```
  if(cervical[i, 30] == TRUE){
    value <- cervical[i, 7] + value
    number <- number + 1
  }
  else{
    value2 <- value2 + cervical[i, 7]
    number2 <- number2 +1
  }
}
mean_cervical_cancer <- value/number
mean_no_cervical_cancer <- value2/number2
mean_cervical_cancer
```

```
## [1] 2.611111
```

```
mean_no_cervical_cancer
```

```
## [1] 2.268371
```

The values displayed show that there is little to no correlation to the amount of pregnancies and whether the subject was diagnosed with cancer

Finally we will examine the correlation between those diagnosed with cancer, and those who were diagnosed with HPV

```
value <- 0
numberofhpv <- 0
value2 <- 0
numberofhpv2 <- 0
#Column 30 is the column indicating whether the subject
#has cancer and column 27 is the column indicating if the subject had hpv or not
for(i in 1:858){
  if(cervical[i, 30] == TRUE){
    if(cervical[i, 29] == TRUE){
      numberofhpv <- numberofhpv + 1
      value <- value +1
    }
    else{
      numberofhpv <- numberofhpv
      value <- value +1
    }
  }
  else{
     if(cervical[i, 29] == TRUE){
      numberofhpv2 <- numberofhpv2 + 1
      value2 <- value2 +1
    }
    else{
      numberofhpv2 <- numberofhpv2
      value2 <- value2 +1
    }
   }
  }
```

```
averagehpvcancer <- numberofhpv/value
averagehpvnocancer <- numberofhpv2/value2
averagehpvcancer
```

```
## [1] 0.1111111
```

```
averagehpvnocancer
```

```
## [1] 0.125
```

The values displayed shows that there is little to no correlation to the diagnosis of hpv and the diagnosis.

Conclusion

Cervical cancer seems to have many factors, those of which can include multiple pregnancies and maybe even age. Given the data that we were provided, we went through several analysis processes and found little to no correlation between any of the major factors of cervical cancer and cervical cancer itself. Now with this data, since it is not from an experiment, we are not scientifically capable of "assuming causation" that is saying "x" causes "y"-but rather we can say that for a general population, we can assume that age, HPV diagnosis and pregnancies do not correlate to the diagnosis of cervical cancer.