Cervical Cancer

Harper Schwab

9/29/2020

Backround

Cervical cancer is the growth of uncontrollable cells that originate in the cervix,which connects the uterus and the vagina. Figure 1 shows a diagram of the female reproductive system.

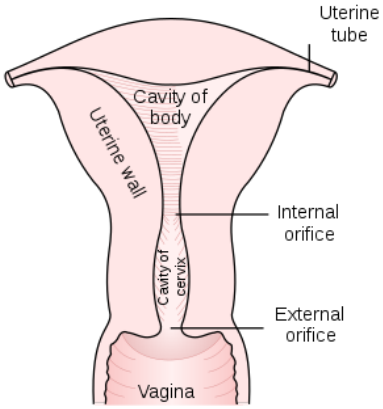


Figure 1

Women are more likely to be diagnosed with cervical cancer after age 30. The greatest cause of cervical cancer is Human Papillomavirus (HPV). HPV, being a sexually transmitted disease, is passed through sexual intercourse and and at least half of sexually active people will have HPV, however the chances for cervical cancer are small.(CDC Basic Information About Cervical Cancer). HPV can have little to no symptoms, and will most likely go away. Other causes for increased risk of cervical cancer is caused by HIV, smoking, extended birth control use, have more than three children, and having multiple sexual partners. (CDC What Are the Risk Factors for Cervical Cancer?) Although the body usually fights HPV before symptoms can appear, the symptom that has a possibility of appearing is genital warts.(Mayo Clinic HPV infection)

There are currently two screening tests available to the public; a Pap test, commonly known as a Pap smear, which looks fro precancel cells; an HPV test which detects the virus. The HPV vaccine can as a preemptive defense against the Virus which is a root cause for cervical cancer. The Vaccine is given to ages 9-26, most commonly given to 11 to 12 year-olds. (CDC What Can I Do to Reduce My Risk of Cervical Cancer?) HPV types 16 and 18 are the cause of 75% of cervical cancer diagnoses types 31 and 45 make up another 10% (Dillman, Robert K.; Oldham, Robert O. 2009). Cervical cancer itself has few symptoms early on. Later in the diseases course abnormal bleeding or discharge from the vagina is the only symptom measurable in the Gynecological Cancer symptom tracker given to patients.

Cervical cancer uses the FIGO system of cancer stages. Stage 1A is when the cancer cells are on the outer lining of the cervix; stage 1B is when the cell mass reaches roughly the size of 4cm; Stage 2A is when the cancer has spread to the top part of the vagina however is localized in the region; Stage 2B is when the cancer has grown grows to areas outside of the cervix into the non-reproductive areas of the body, however is still within the general are of the cervix; Stage 3B is where the cancer cells are spreading to the kidneys and have blocked off the uterus; 4A occurs when the cancer cells spread to areas such as the womb, bladder, vagina, and rectum;Stage 4B cervical cancer is when the cancer has spread to the lungs. (Cervical cancer Wikipedia)

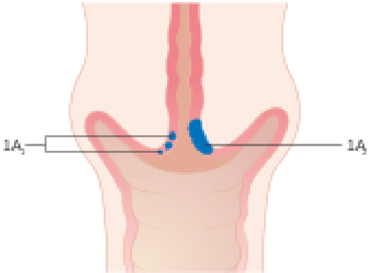


Figure 1

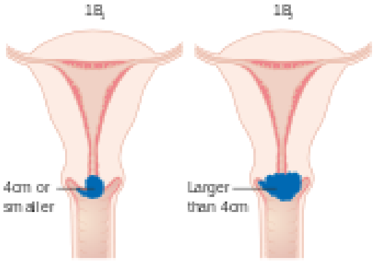


Figure 1

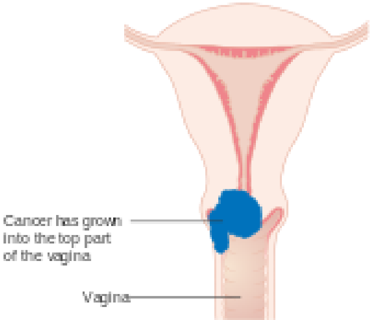


Figure 1

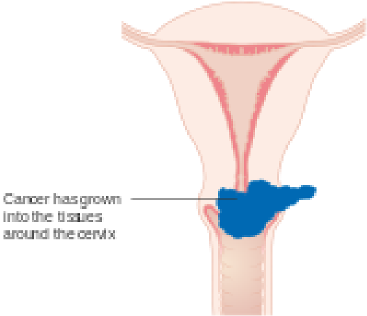


Figure 1

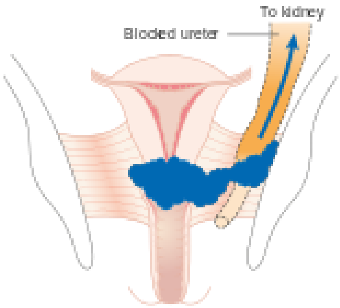


Figure 1

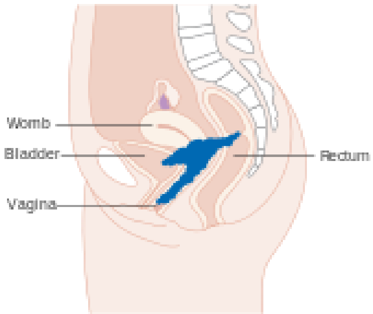


Figure 1

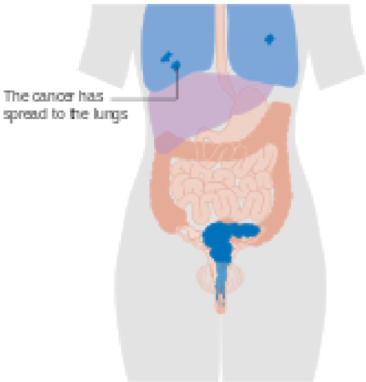


Figure 1

Cervical Cancer has many possible risk factors, personally, I am most interested to look in to how smoking, amount of sexual partners, taking hormonal contraceptives for longer than 5 years,and HPV affect cervical cancer.

Importing data

This data is a cleaned form of original data from the UCI Machine Learning Repository, which was collected in Caracas Venezuela at the Hospital Universitario de Caracas.

```
summary(data)
```

```
##       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
##  num_of_pregnancies   smokes         smokes_years   smokes_packs_year
##  Min.   : 0.000     Mode :logical   Min.   : 0.00   Min.   : 0.0000
##  1st Qu.: 1.000     FALSE:735       1st Qu.: 0.00   1st Qu.: 0.0000
##  Median : 2.000     TRUE :123       Median : 0.00   Median : 0.0000
##  Mean   : 2.276                     Mean   : 1.22   Mean   : 0.4531
##  3rd Qu.: 3.000                     3rd Qu.: 0.00   3rd Qu.: 0.0000
##  Max.   :11.000                     Max.   :37.00   Max.   :37.0000
##  hormonal_contraceptives hormonal_contraceptives_years    iud
##  Mode :logical           Min.   : 0.000                Mode :logical
##  FALSE:377               1st Qu.: 0.000                FALSE:775
##  TRUE :481               Median : 1.000                TRUE :83
##                          Mean   : 2.256
##                          3rd Qu.: 2.256
##                          Max.   :30.000
##    iud_years          stds          stds_number     stds_condylomatosis
##  Min.   : 0.0000   Mode :logical   Min.   :0.0000   Mode :logical
##  1st Qu.: 0.0000   FALSE:779       1st Qu.:0.0000   FALSE:814
##  Median : 0.0000   TRUE :79        Median :0.0000   TRUE :44
##  Mean   : 0.5148                   Mean   :0.1766
##  3rd Qu.: 0.0000                   3rd Qu.:0.0000
##  Max.   :19.0000                   Max.   :4.0000
##  stds_cervical_condylomatosis stds_vaginal_condylomatosis
##  Mode :logical                Mode :logical
##  FALSE:858                    FALSE:854
##                               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 stds_time_since_first_diagnosis
##  Mode :logical   Min.   :0.00000          Min.   : 1.000
##  FALSE:856       1st Qu.:0.00000          1st Qu.: 6.141
##  TRUE :2         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 dx_cancer         dx_cin           dx_hpv
##  Min.   : 1.000                 Mode :logical   Mode :logical   Mode :logical
##  1st Qu.: 5.817                 FALSE:840       FALSE:849       FALSE:840
##  Median : 5.817                 TRUE :18        TRUE :9         TRUE :18
##  Mean   : 5.817
##  3rd Qu.: 5.817
##  Max.   :22.000
##      dx          hinselmann       schiller        citology
##  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
##
##
##
```

With the summary command we can see an overview of some distriptave statistics. Later, we will look at them more closely.

```
head(data)
```

```
##   age number_of_sexual_partners first_sexual_intercourse num_of_pregnancies
## 1  18                         4                  15.0000                  1
## 2  15                         1                  14.0000                  1
## 3  34                         1                  16.9953                  1
## 4  52                         5                  16.0000                  4
## 5  46                         3                  21.0000                  4
## 6  42                         3                  23.0000                  2
##   smokes smokes_years smokes_packs_year hormonal_contraceptives
## 1  FALSE            0                 0                   FALSE
## 2  FALSE            0                 0                   FALSE
## 3  FALSE            0                 0                   FALSE
## 4   TRUE           37                37                    TRUE
## 5  FALSE            0                 0                    TRUE
## 6  FALSE            0                 0                   FALSE
##   hormonal_contraceptives_years   iud iud_years  stds stds_number
## 1                             0 FALSE         0 FALSE           0
## 2                             0 FALSE         0 FALSE           0
## 3                             0 FALSE         0 FALSE           0
## 4                             3 FALSE         0 FALSE           0
## 5                            15 FALSE         0 FALSE           0
## 6                             0 FALSE         0 FALSE           0
##   stds_condylomatosis stds_cervical_condylomatosis stds_vaginal_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_vulvo_perineal_condylomatosis stds_syphilis
## 1                              FALSE         FALSE
## 2                              FALSE         FALSE
## 3                              FALSE         FALSE
## 4                              FALSE         FALSE
## 5                              FALSE         FALSE
## 6                              FALSE         FALSE
##   stds_pelvic_inflammatory_disease stds_genital_herpes
## 1                            FALSE               FALSE
## 2                            FALSE               FALSE
## 3                            FALSE               FALSE
## 4                            FALSE               FALSE
## 5                            FALSE               FALSE
## 6                            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
##   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 dx_cancer dx_cin dx_hpv    dx hinselmann
## 1                       5.816901     FALSE  FALSE  FALSE FALSE      FALSE
## 2                       5.816901     FALSE  FALSE  FALSE FALSE      FALSE
## 3                       5.816901     FALSE  FALSE  FALSE FALSE      FALSE
## 4                       5.816901      TRUE  FALSE   TRUE FALSE      FALSE
## 5                       5.816901     FALSE  FALSE  FALSE FALSE      FALSE
## 6                       5.816901     FALSE  FALSE  FALSE FALSE      FALSE
##   schiller citology biopsy
## 1    FALSE    FALSE  FALSE
## 2    FALSE    FALSE  FALSE
## 3    FALSE    FALSE  FALSE
## 4    FALSE    FALSE  FALSE
## 5    FALSE    FALSE  FALSE
## 6    FALSE    FALSE  FALSE
```

```
tail(data)
```

```
##     age number_of_sexual_partners first_sexual_intercourse num_of_pregnancies
## 853  43                         3                       17                  3
## 854  34                         3                       18                  0
## 855  32                         2                       19                  1
## 856  25                         2                       17                  0
## 857  33                         2                       24                  2
## 858  29                         2                       20                  1
##     smokes smokes_years smokes_packs_year hormonal_contraceptives
## 853  FALSE            0                 0                    TRUE
## 854  FALSE            0                 0                   FALSE
## 855  FALSE            0                 0                    TRUE
## 856  FALSE            0                 0                    TRUE
## 857  FALSE            0                 0                    TRUE
## 858  FALSE            0                 0                    TRUE
##     hormonal_contraceptives_years   iud iud_years  stds stds_number
## 853                          5.00 FALSE         0 FALSE           0
## 854                          0.00 FALSE         0 FALSE           0
## 855                          8.00 FALSE         0 FALSE           0
## 856                          0.08 FALSE         0 FALSE           0
## 857                          0.08 FALSE         0 FALSE           0
## 858                          0.50 FALSE         0 FALSE           0
##     stds_condylomatosis stds_cervical_condylomatosis
## 853               FALSE                        FALSE
## 854               FALSE                        FALSE
## 855               FALSE                        FALSE
## 856               FALSE                        FALSE
## 857               FALSE                        FALSE
## 858               FALSE                        FALSE
##     stds_vaginal_condylomatosis stds_vulvo_perineal_condylomatosis
## 853                       FALSE                              FALSE
## 854                       FALSE                              FALSE
## 855                       FALSE                              FALSE
## 856                       FALSE                              FALSE
## 857                       FALSE                              FALSE
## 858                       FALSE                              FALSE
##     stds_syphilis stds_pelvic_inflammatory_disease stds_genital_herpes
## 853         FALSE                            FALSE               FALSE
## 854         FALSE                            FALSE               FALSE
## 855         FALSE                            FALSE               FALSE
## 856         FALSE                            FALSE               FALSE
## 857         FALSE                            FALSE               FALSE
## 858         FALSE                            FALSE               FALSE
##     stds_molluscum_contagiosum stds_aids stds_hiv stds_hepatitis_b stds_hpv
## 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
##     stds_number_of_diagnosis stds_time_since_first_diagnosis
## 853                        0                        6.140845
## 854                        0                        6.140845
## 855                        0                        6.140845
## 856                        0                        6.140845
## 857                        0                        6.140845
## 858                        0                        6.140845
##     stds_time_since_last_diagnosis dx_cancer dx_cin dx_hpv    dx hinselmann
## 853                       5.816901     FALSE  FALSE  FALSE FALSE      FALSE
## 854                       5.816901     FALSE  FALSE  FALSE FALSE      FALSE
## 855                       5.816901     FALSE  FALSE  FALSE FALSE      FALSE
## 856                       5.816901     FALSE  FALSE  FALSE FALSE      FALSE
## 857                       5.816901     FALSE  FALSE  FALSE FALSE      FALSE
## 858                       5.816901     FALSE  FALSE  FALSE FALSE      FALSE
##     schiller citology biopsy
## 853    FALSE    FALSE  FALSE
## 854    FALSE    FALSE  FALSE
## 855    FALSE    FALSE  FALSE
## 856    FALSE     TRUE  FALSE
## 857    FALSE    FALSE  FALSE
## 858    FALSE    FALSE  FALSE
```

Looking at the head and the tail of the data help us diagnose if there were any errors reading and importing the data. It seems that there haven’t been any, so I was able to continue the analysis.

```
glimpse(data)
```

```
## Rows: 858
## Columns: 36
## $ age                                <int> 18, 15, 34, 52, 46, 42, 51, 26, ...
## $ number_of_sexual_partners          <dbl> 4, 1, 1, 5, 3, 3, 3, 1, 1, 3, 3,...
## $ first_sexual_intercourse           <dbl> 15.0000, 14.0000, 16.9953, 16.00...
## $ num_of_pregnancies                 <dbl> 1.000000, 1.000000, 1.000000, 4....
## $ smokes                             <lgl> FALSE, FALSE, FALSE, TRUE, FALSE...
## $ smokes_years                       <dbl> 0.000000, 0.000000, 0.000000, 37...
## $ smokes_packs_year                  <dbl> 0.0, 0.0, 0.0, 37.0, 0.0, 0.0, 3...
## $ hormonal_contraceptives            <lgl> FALSE, FALSE, FALSE, TRUE, TRUE,...
## $ hormonal_contraceptives_years      <dbl> 0.00, 0.00, 0.00, 3.00, 15.00, 0...
## $ iud                                <lgl> FALSE, FALSE, FALSE, FALSE, FALS...
## $ iud_years                          <dbl> 0.0000000, 0.0000000, 0.0000000,...
## $ stds                               <lgl> FALSE, FALSE, FALSE, FALSE, FALS...
## $ stds_number                        <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ stds_condylomatosis                <lgl> FALSE, FALSE, FALSE, FALSE, FALS...
## $ stds_cervical_condylomatosis       <lgl> FALSE, FALSE, FALSE, FALSE, FALS...
## $ stds_vaginal_condylomatosis        <lgl> FALSE, FALSE, FALSE, FALSE, FALS...
## $ stds_vulvo_perineal_condylomatosis <lgl> FALSE, FALSE, FALSE, FALSE, FALS...
## $ stds_syphilis                      <lgl> FALSE, FALSE, FALSE, FALSE, FALS...
## $ stds_pelvic_inflammatory_disease   <lgl> FALSE, FALSE, FALSE, FALSE, FALS...
## $ stds_genital_herpes                <lgl> FALSE, FALSE, FALSE, FALSE, FALS...
## $ stds_molluscum_contagiosum         <lgl> FALSE, FALSE, FALSE, FALSE, FALS...
## $ stds_aids                          <lgl> FALSE, FALSE, FALSE, FALSE, FALS...
## $ stds_hiv                           <lgl> FALSE, FALSE, FALSE, FALSE, FALS...
## $ stds_hepatitis_b                   <lgl> FALSE, FALSE, FALSE, FALSE, FALS...
## $ stds_hpv                           <lgl> FALSE, FALSE, FALSE, FALSE, FALS...
## $ stds_number_of_diagnosis           <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ stds_time_since_first_diagnosis    <dbl> 6.140845, 6.140845, 6.140845, 6....
## $ stds_time_since_last_diagnosis     <dbl> 5.816901, 5.816901, 5.816901, 5....
## $ dx_cancer                          <lgl> FALSE, FALSE, FALSE, TRUE, FALSE...
## $ dx_cin                             <lgl> FALSE, FALSE, FALSE, FALSE, FALS...
## $ dx_hpv                             <lgl> FALSE, FALSE, FALSE, TRUE, FALSE...
## $ dx                                 <lgl> FALSE, FALSE, FALSE, FALSE, FALS...
## $ hinselmann                         <lgl> FALSE, FALSE, FALSE, FALSE, FALS...
## $ schiller                           <lgl> FALSE, FALSE, FALSE, FALSE, FALS...
## $ citology                           <lgl> FALSE, FALSE, FALSE, FALSE, FALS...
## $ biopsy                             <lgl> FALSE, FALSE, FALSE, FALSE, FALS...
```

We can easily see the different types of data. Most of the data available are logicals (True/False). Look to the BIC modeling section to see more done with these categories.

```
names(data)
```

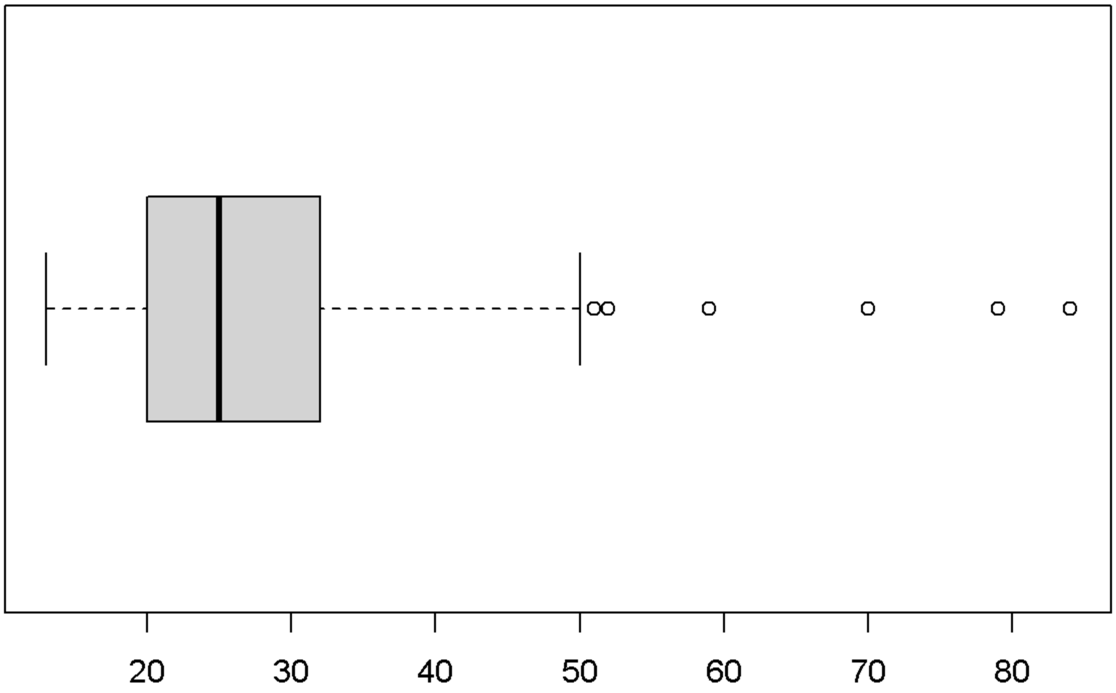
```
##  [1] "age"                                "number_of_sexual_partners"
##  [3] "first_sexual_intercourse"           "num_of_pregnancies"
##  [5] "smokes"                             "smokes_years"
##  [7] "smokes_packs_year"                  "hormonal_contraceptives"
##  [9] "hormonal_contraceptives_years"      "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"
```

For future reference it is always good to know what the names of each category you use is titled. In this case the names() command helps us with that.

Plots and Analysis

One thing that is important to consider the ages of the participants. Cervical cancer is more likely in poeple who are older in age, therefore we must take that into account in the results of our data.

```
boxplot(data$age, horizontal = TRUE)
```



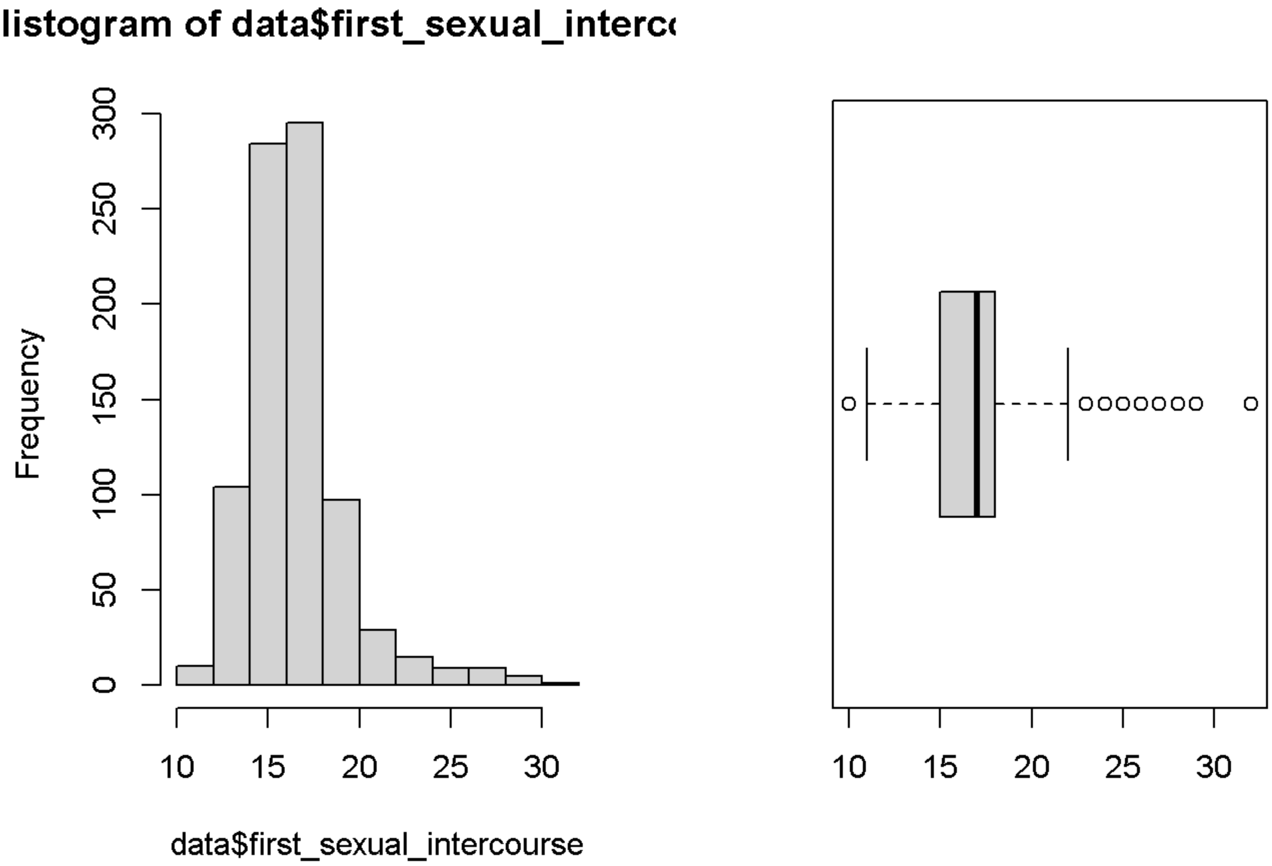
```
mean(data$age)
```

```
## [1] 26.82051
```

We can see that there is a strong skew left, towards younger people in our sample. if we had more of a even spread, and mean closer to 30 and older then the data can be more accurate to the whole population so we must remember that exception to our results. The average age of those surveyed is roughly 27 years old.

Another factor in diagnoses how early the patient became sexually active.

```
par(mfrow=c(1,2))
hist(data$first_sexual_intercourse)
boxplot(data$first_sexual_intercourse, horizontal = TRUE)
```

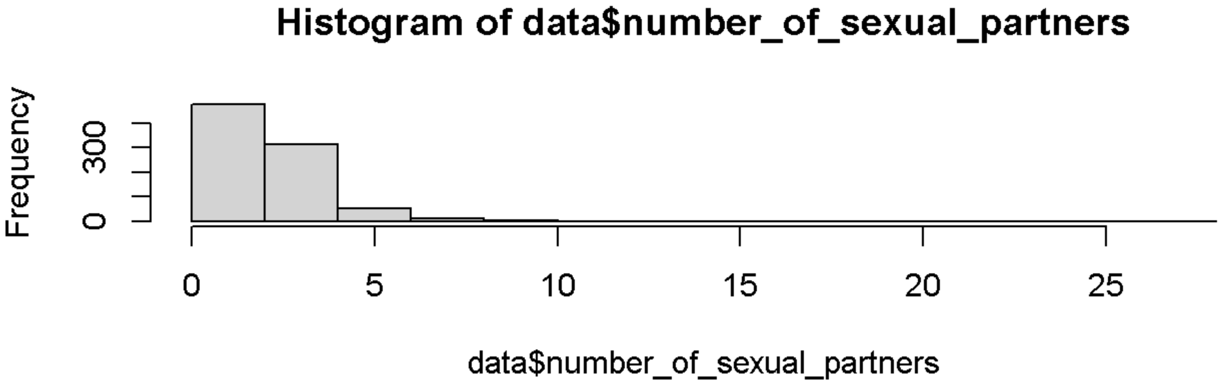


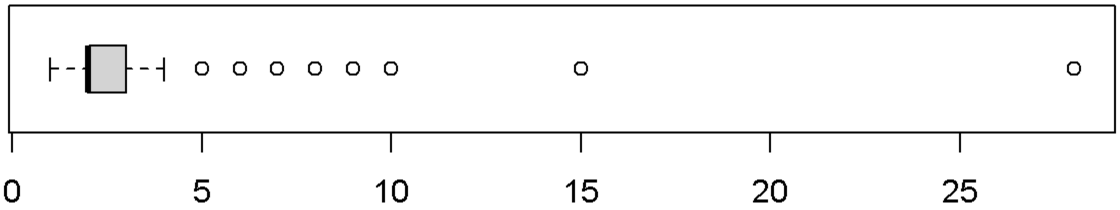
```
par(mfrow=c(1,1))
```

Most people had sex for the first time between ages 15 and 19. With an average at age 16.9953, so almost 17 years of age. We can start to consider people below the 25th percentile early, because the middle 50% can be considered the average.

Next I will look at the number of sexual partners in a histogram.

```
par(mfrow=c(2,1))
hist(data$number_of_sexual_partners)
boxplot(data$number_of_sexual_partners, horizontal = TRUE)
```





```
par(mfrow=c(1,1))
mean(data$number_of_sexual_partners)
```

```
## [1] 2.527644
```

We can see the majority of people have less that 4 sexual partners, the mean of number of sexual partners is 2.527644. So on average, people have between 2-3 sexual partners when they are around 27 years old.

For smokers I wanted to be able to single them out, so I filtered the data to only include smokers.

```
smokes <- filter(data, data$smokes == TRUE)
head(smokes)
```

```
##   age number_of_sexual_partners first_sexual_intercourse num_of_pregnancies
## 1  52                         5                       16           4.000000
## 2  51                         3                       17           6.000000
## 3  44                         3                       15           2.275561
## 4  37                         3                       24           1.000000
## 5  37                         3                       17           5.000000
## 6  36                         3                       18           3.000000
##   smokes smokes_years smokes_packs_year hormonal_contraceptives
## 1   TRUE    37.000000        37.0000000                    TRUE
## 2   TRUE    34.000000         3.4000000                   FALSE
## 3   TRUE     1.266973         2.8000000                   FALSE
## 4   TRUE     3.000000         0.0400000                   FALSE
## 5   TRUE     1.266973         0.5132021                    TRUE
## 6   TRUE     1.266973         2.4000000                    TRUE
##   hormonal_contraceptives_years   iud iud_years  stds stds_number
## 1                             3 FALSE 0.0000000 FALSE           0
## 2                             0  TRUE 7.0000000 FALSE           0
## 3                             0 FALSE 0.5148043 FALSE           0
## 4                             0 FALSE 0.0000000 FALSE           0
## 5                            10 FALSE 0.0000000  TRUE           1
## 6                             9 FALSE 0.0000000 FALSE           0
##   stds_condylomatosis stds_cervical_condylomatosis stds_vaginal_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_vulvo_perineal_condylomatosis stds_syphilis
## 1                              FALSE         FALSE
## 2                              FALSE         FALSE
## 3                              FALSE         FALSE
## 4                              FALSE         FALSE
## 5                              FALSE          TRUE
## 6                              FALSE         FALSE
##   stds_pelvic_inflammatory_disease stds_genital_herpes
## 1                            FALSE               FALSE
## 2                            FALSE               FALSE
## 3                            FALSE               FALSE
## 4                            FALSE               FALSE
## 5                            FALSE               FALSE
## 6                            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
##   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 dx_cancer dx_cin dx_hpv    dx hinselmann
## 1                       5.816901      TRUE  FALSE   TRUE FALSE      FALSE
## 2                       5.816901     FALSE  FALSE  FALSE FALSE       TRUE
## 3                       5.816901     FALSE  FALSE  FALSE FALSE      FALSE
## 4                       5.816901     FALSE  FALSE  FALSE FALSE      FALSE
## 5                       5.816901     FALSE  FALSE  FALSE FALSE      FALSE
## 6                       5.816901     FALSE  FALSE  FALSE FALSE      FALSE
##   schiller citology biopsy
## 1    FALSE    FALSE  FALSE
## 2     TRUE    FALSE   TRUE
## 3    FALSE    FALSE  FALSE
## 4    FALSE    FALSE  FALSE
## 5    FALSE    FALSE  FALSE
## 6    FALSE    FALSE  FALSE
```

More specifically I wanted to see smokers who also have STDs

```
smokeSTD <- filter(data, data$smokes == TRUE,
                 data$stds == TRUE)
head(smokeSTD)
```

```
##   age number_of_sexual_partners first_sexual_intercourse num_of_pregnancies
## 1  37                  3.000000                       17                  5
## 2  35                  3.000000                       17                  6
## 3  30                  2.527644                       13                  3
## 4  28                  3.000000                       16                  3
## 5  27                  2.000000                       13                  2
## 6  26                  3.000000                       15                  3
##   smokes smokes_years smokes_packs_year hormonal_contraceptives
## 1   TRUE     1.266973         0.5132021                    TRUE
## 2   TRUE    13.000000         2.6000000                    TRUE
## 3   TRUE    22.000000         3.3000000                   FALSE
## 4   TRUE    12.000000         6.0000000                    TRUE
## 5   TRUE     7.000000         1.4000000                    TRUE
## 6   TRUE     5.000000         5.0000000                   FALSE
##   hormonal_contraceptives_years   iud iud_years stds stds_number
## 1                     10.000000 FALSE 0.0000000 TRUE           1
## 2                      7.000000 FALSE 0.0000000 TRUE           1
## 3                      0.000000 FALSE 0.0000000 TRUE           1
## 4                      7.000000 FALSE 0.0000000 TRUE           1
## 5                      3.000000 FALSE 0.0000000 TRUE           3
## 6                      2.256419 FALSE 0.5148043 TRUE           1
##   stds_condylomatosis stds_cervical_condylomatosis stds_vaginal_condylomatosis
## 1               FALSE                        FALSE                       FALSE
## 2               FALSE                        FALSE                       FALSE
## 3               FALSE                        FALSE                       FALSE
## 4               FALSE                        FALSE                       FALSE
## 5                TRUE                        FALSE                       FALSE
## 6               FALSE                        FALSE                       FALSE
##   stds_vulvo_perineal_condylomatosis stds_syphilis
## 1                              FALSE          TRUE
## 2                              FALSE          TRUE
## 3                              FALSE         FALSE
## 4                              FALSE         FALSE
## 5                               TRUE          TRUE
## 6                              FALSE         FALSE
##   stds_pelvic_inflammatory_disease stds_genital_herpes
## 1                            FALSE               FALSE
## 2                            FALSE               FALSE
## 3                            FALSE               FALSE
## 4                            FALSE               FALSE
## 5                            FALSE               FALSE
## 6                            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     TRUE            FALSE    FALSE
## 4                      FALSE     FALSE     TRUE            FALSE    FALSE
## 5                      FALSE     FALSE    FALSE            FALSE    FALSE
## 6                      FALSE     FALSE     TRUE            FALSE    FALSE
##   stds_number_of_diagnosis stds_time_since_first_diagnosis
## 1                        0                        6.140845
## 2                        1                       12.000000
## 3                        1                        3.000000
## 4                        1                        2.000000
## 5                        1                        5.000000
## 6                        1                        6.000000
##   stds_time_since_last_diagnosis dx_cancer dx_cin dx_hpv    dx hinselmann
## 1                       5.816901     FALSE  FALSE  FALSE FALSE      FALSE
## 2                      12.000000     FALSE  FALSE  FALSE FALSE      FALSE
## 3                       3.000000     FALSE  FALSE  FALSE FALSE       TRUE
## 4                       2.000000     FALSE  FALSE  FALSE FALSE      FALSE
## 5                       5.000000     FALSE  FALSE  FALSE FALSE      FALSE
## 6                       6.000000     FALSE  FALSE  FALSE FALSE      FALSE
##   schiller citology biopsy
## 1    FALSE    FALSE  FALSE
## 2     TRUE    FALSE  FALSE
## 3     TRUE    FALSE   TRUE
## 4    FALSE    FALSE  FALSE
## 5    FALSE    FALSE  FALSE
## 6    FALSE    FALSE  FALSE
```

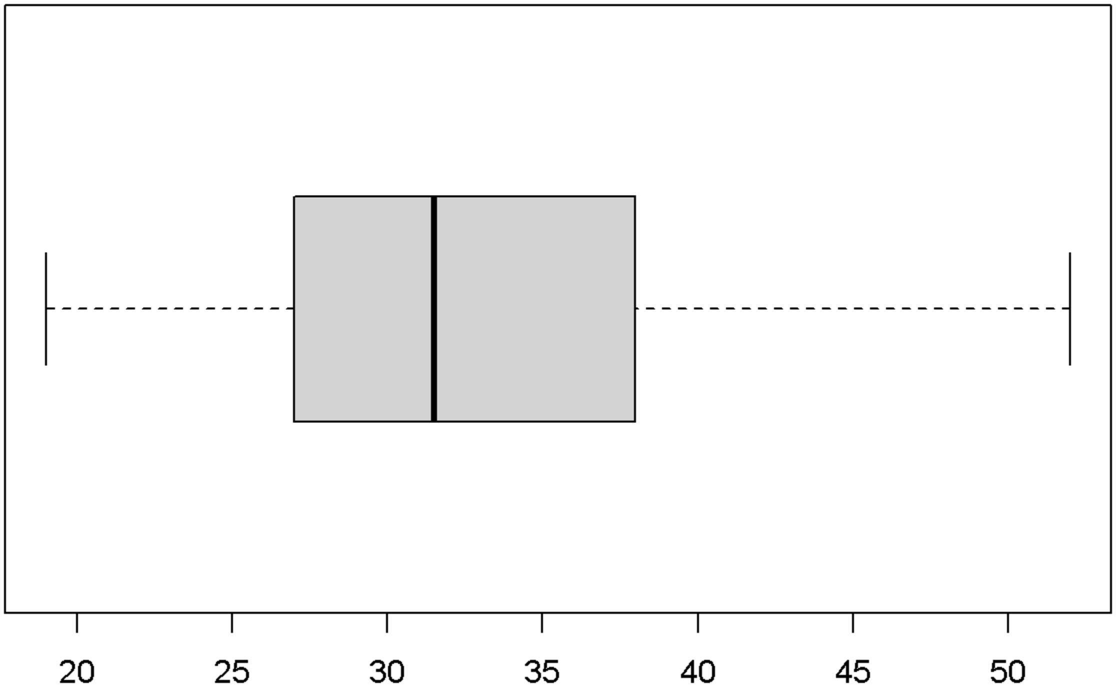
Surprisingly, nobody in that category has cervical cancer and only one of them has HPV. So far it seems like smoking has less of an affect on weather or not you might have cervical cancer then I thought previously.

Then I wanted to take a look at all people who actually had cervical cancer to see if I could see anything just by looking at the data, before deeper analysis.

```
Yescancer <- filter(data, data$dx_cancer == TRUE)
```

First, the best thing is to look at the ages of these 18 individuals.

```
boxplot(Yescancer$age, horizontal = TRUE)
```



```
quantile(Yescancer$age)
```

```
##   0%  25%  50%  75% 100%
## 19.0 27.5 31.5 38.0 52.0
```

```
mean(Yescancer$age)
```

```
## [1] 33.22222
```

We can tell that most of the people who have cervical cancer are between ages 27 and 38 with the average being 33. This could just be because our sample is skewed towards a younger demographic.

Another interesting thing is that 13 out of the 18 people with cervical cancer take hormonal contraceptives, and the only STD any of them had was HPV, and only one of them had it. So we can infer from just the results that hormonal contraceptives contribute to cervical cancer.

To try to see how HPV is spread in the data I created a dataset with only people with HPV.

```
hpv <- filter(data, data$stds_hpv == TRUE)
head(hpv)
```

```
##   age number_of_sexual_partners first_sexual_intercourse num_of_pregnancies
## 1  36                         3                       20                  2
## 2  32                         3                       18                  1
##   smokes smokes_years smokes_packs_year hormonal_contraceptives
## 1  FALSE            0              0.00                    TRUE
## 2   TRUE           11              0.16                    TRUE
##   hormonal_contraceptives_years   iud iud_years stds stds_number
## 1                             6 FALSE         0 TRUE           1
## 2                             6 FALSE         0 TRUE           1
##   stds_condylomatosis stds_cervical_condylomatosis stds_vaginal_condylomatosis
## 1               FALSE                        FALSE                       FALSE
## 2               FALSE                        FALSE                       FALSE
##   stds_vulvo_perineal_condylomatosis stds_syphilis
## 1                              FALSE         FALSE
## 2                              FALSE         FALSE
##   stds_pelvic_inflammatory_disease stds_genital_herpes
## 1                            FALSE               FALSE
## 2                            FALSE               FALSE
##   stds_molluscum_contagiosum stds_aids stds_hiv stds_hepatitis_b stds_hpv
## 1                      FALSE     FALSE    FALSE            FALSE     TRUE
## 2                      FALSE     FALSE    FALSE            FALSE     TRUE
##   stds_number_of_diagnosis stds_time_since_first_diagnosis
## 1                        1                       16.000000
## 2                        0                        6.140845
##   stds_time_since_last_diagnosis dx_cancer dx_cin dx_hpv    dx hinselmann
## 1                      16.000000      TRUE  FALSE   TRUE  TRUE      FALSE
## 2                       5.816901      TRUE  FALSE   TRUE FALSE      FALSE
##   schiller citology biopsy
## 1    FALSE    FALSE  FALSE
## 2    FALSE    FALSE  FALSE
```

The only two people who had HPV in the study show that HPV is rare and, based on the data, HPV can give a 50% chance of cervical cancer. However, HPV many times, is fought off without the patient knowing they had it. So this number while showing the possibility of HPV giving a 50% chance of cancer, does not give enough data to back up that claim.

To look at hormonal contraceptives I decided to create a data set that only has those who take the contraceptives and for longer than 5 years.

```
cont5<- filter(data, data$hormonal_contraceptives == TRUE,
               data$hormonal_contraceptives_years >4)
head(cont5)
```

```
##   age number_of_sexual_partners first_sexual_intercourse num_of_pregnancies
## 1  46                         3                       21           4.000000
## 2  27                         1                       17           3.000000
## 3  45                         4                       14           6.000000
## 4  44                         2                       25           2.000000
## 5  40                         3                       18           2.000000
## 6  42                         2                       20           2.275561
##   smokes smokes_years smokes_packs_year hormonal_contraceptives
## 1  FALSE            0                 0                    TRUE
## 2  FALSE            0                 0                    TRUE
## 3  FALSE            0                 0                    TRUE
## 4  FALSE            0                 0                    TRUE
## 5  FALSE            0                 0                    TRUE
## 6  FALSE            0                 0                    TRUE
##   hormonal_contraceptives_years   iud iud_years  stds stds_number
## 1                            15 FALSE         0 FALSE           0
## 2                             8 FALSE         0 FALSE           0
## 3                            10  TRUE         5 FALSE           0
## 4                             5 FALSE         0 FALSE           0
## 5                            15 FALSE         0 FALSE           0
## 6                             7  TRUE         6  TRUE           2
##   stds_condylomatosis stds_cervical_condylomatosis stds_vaginal_condylomatosis
## 1               FALSE                        FALSE                       FALSE
## 2               FALSE                        FALSE                       FALSE
## 3               FALSE                        FALSE                       FALSE
## 4               FALSE                        FALSE                       FALSE
## 5               FALSE                        FALSE                       FALSE
## 6                TRUE                        FALSE                       FALSE
##   stds_vulvo_perineal_condylomatosis stds_syphilis
## 1                              FALSE         FALSE
## 2                              FALSE         FALSE
## 3                              FALSE         FALSE
## 4                              FALSE         FALSE
## 5                              FALSE         FALSE
## 6                               TRUE         FALSE
##   stds_pelvic_inflammatory_disease stds_genital_herpes
## 1                            FALSE               FALSE
## 2                            FALSE               FALSE
## 3                            FALSE               FALSE
## 4                            FALSE               FALSE
## 5                            FALSE               FALSE
## 6                            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
##   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 dx_cancer dx_cin dx_hpv    dx hinselmann
## 1                       5.816901     FALSE  FALSE  FALSE FALSE      FALSE
## 2                       5.816901     FALSE  FALSE  FALSE FALSE      FALSE
## 3                       5.816901     FALSE  FALSE  FALSE FALSE      FALSE
## 4                       5.816901     FALSE  FALSE  FALSE FALSE      FALSE
## 5                       5.816901     FALSE  FALSE  FALSE FALSE      FALSE
## 6                       5.816901     FALSE  FALSE  FALSE FALSE      FALSE
##   schiller citology biopsy
## 1    FALSE    FALSE  FALSE
## 2    FALSE    FALSE  FALSE
## 3    FALSE    FALSE  FALSE
## 4    FALSE    FALSE  FALSE
## 5    FALSE    FALSE  FALSE
## 6    FALSE    FALSE  FALSE
```

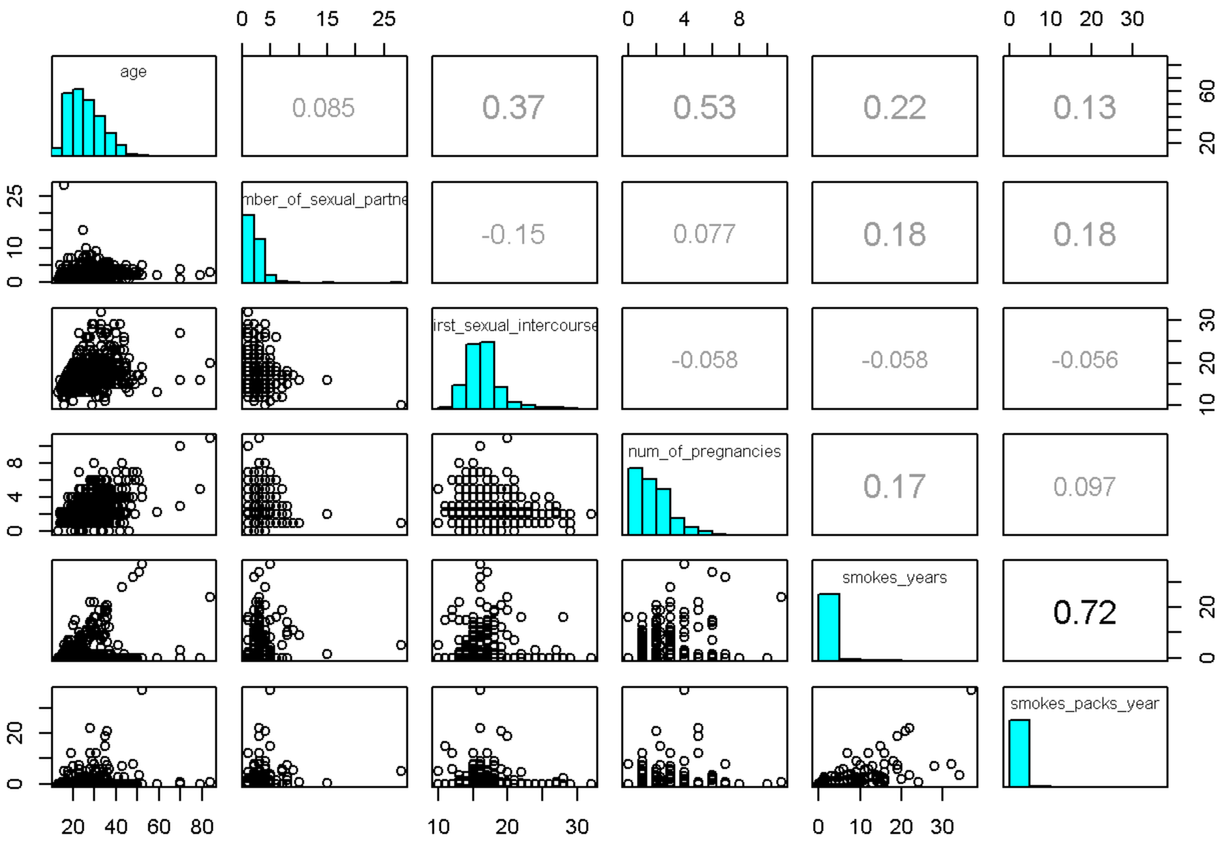
I wanted to narrow it down further so I added the parameter to have cancer.

```
cont5can<- filter(data, data$hormonal_contraceptives == TRUE,
               data$hormonal_contraceptives_years >4,
               data$dx_cancer == TRUE)
```

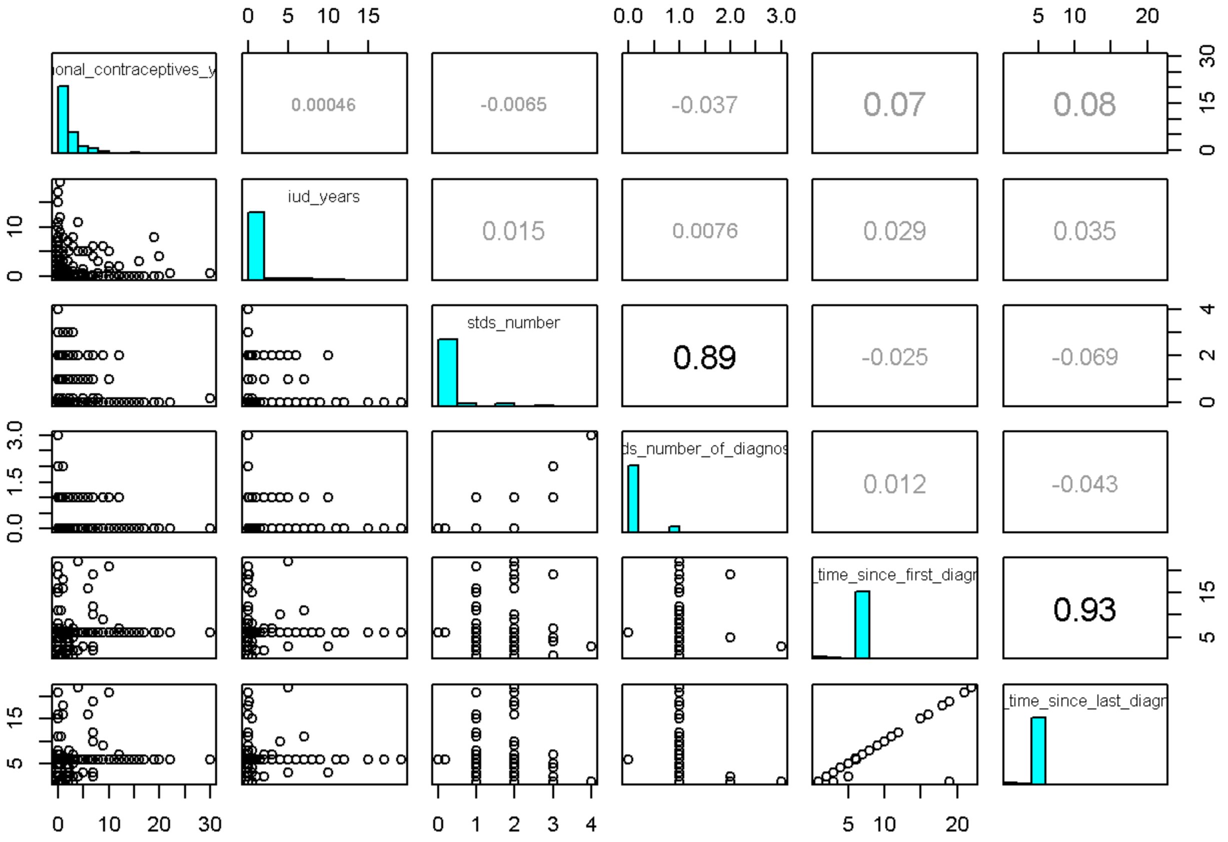
Two out of the five individuals have HPV as well as have taken hormonal contraceptives for over 5 years. It seems like the contraceptives are a important say in the chances of developing ce3rvix cancer. But multiple factors defiantly add on to it, like HPV.

To see some overall correlations I did a few pairwise plots. All of the numeric data is a little to big to put onto one, so I split it in half.

```
numerics <- data[, c(1,2,3,4,6,7,9,11,13,26,27,28)]
pairs(numerics[1:6],upper.panel = panel.cor,diag.panel=panel.hist)
```



```
pairs(numerics[7:12],upper.panel = panel.cor,diag.panel=panel.hist)
```



Take a look at the different correlations, some obvious ones jump out pretty fast. For instance the higher the number of sexual partners, the younger you were when you first had sexual intercourse.

BIC Modeling

BIC modeling allows for important causal studies of different factors. One of the first factors I want to test is smoking to cancer.

```
#BIC analysis
#Does smoking have an affect on Cancer?
cansmoke<-lm(data$dx_cancer~data$smokes) #Comparing here to look at the summary
summary(cansmoke)
```

```
## 
## Call:
## lm(formula = data$dx_cancer ~ data$smokes)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.02177 -0.02177 -0.02177 -0.02177  0.98374 
## 
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)    
## (Intercept)      0.021769   0.005292   4.114 4.27e-05 ***
## data$smokesTRUE -0.005509   0.013977  -0.394    0.694    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.1435 on 856 degrees of freedom
## Multiple R-squared:  0.0001814,  Adjusted R-squared:  -0.0009866 
## F-statistic: 0.1553 on 1 and 856 DF,  p-value: 0.6936
```

```
BIC(lm(data$dx_cancer~1)) #Comparing to one
```

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

```
BIC(lm(data$dx_cancer~data$smokes)) #looking at difference
```

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

```
#slight
```

We can tell from the BIC analysis that smoking does have an affect on cervix cancer, however, it does is not a major affect because difference is less than 10.

Another factor I wanted to look at the amount of sexual partners the subjects had.

```
#BIC analysis
#Does Sexual partner #  have an affect on Cancer?
canpart<-lm(data$dx_cancer~data$number_of_sexual_partners) #Comparing here to look at the summary
summary(canpart)
```

```
## 
## Call:
## lm(formula = data$dx_cancer ~ data$number_of_sexual_partners)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.07060 -0.02190 -0.01995 -0.01800  0.98200 
## 
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)  
## (Intercept)                     0.016055   0.008992   1.785   0.0745 .
## data$number_of_sexual_partners 0.001948   0.002984   0.653   0.5140  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.1434 on 856 degrees of freedom
## Multiple R-squared:  0.0004977,  Adjusted R-squared:  -0.00067 
## F-statistic: 0.4262 on 1 and 856 DF,  p-value: 0.514
```

```
BIC(lm(data$dx_cancer~1)) #Comparing to one
```

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

```
BIC(lm(data$dx_cancer~data$number_of_sexual_partners)) #looking at difference
```

```
## [1] -878.9676
```

```
#slight
```

Again, much like smoking the amount of sexual partners an individual has affects cervical cancer slightly but not majorly. Another factor I wanted to look at the amount of sexual partners the subjects had.

```
#BIC analysis
#do Hormonal contracepticves  have an affect on Cancer?
canhor<-lm(data$dx_cancer~data$hormonal_contraceptives) #Comparing here to look at the summary
summary(canhor)
```

```
##
## Call:
## lm(formula = data$dx_cancer ~ data$hormonal_contraceptives)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.02703 -0.02703 -0.02703 -0.01326  0.98674
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      0.013263   0.007381   1.797   0.0727 .
## data$hormonal_contraceptivesTRUE 0.013764   0.009858   1.396   0.1630
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1433 on 856 degrees of freedom
## Multiple R-squared:  0.002272,   Adjusted R-squared:  0.001107
## F-statistic: 1.949 on 1 and 856 DF,  p-value: 0.163
```

```
BIC(lm(data$dx_cancer~1)) #Comparing to one
```

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

```
BIC(lm(data$dx_cancer~data$hormonal_contraceptives)) #looking at difference
```

```
## [1] -880.4923
```

```
#slight
```

By itself, just taking hormonal contraceptives does not have a significant affect on cervical cancer. Next thing to look at is HPV and Cancer.

```
#BIC analysis
#Does HPV  have an affect on Cancer?
canhor<-lm(data$dx_cancer~data$stds_hpv) #Comparing here to look at the summary
summary(canhor)
```

```
##
## Call:
## lm(formula = data$dx_cancer ~ data$stds_hpv)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.01869 -0.01869 -0.01869 -0.01869  0.98131
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.018692   0.004629   4.038 5.88e-05 ***
## data$stds_hpvTRUE 0.981308   0.095878  10.235  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1354 on 856 degrees of freedom
## Multiple R-squared:  0.109,  Adjusted R-squared:  0.108
## F-statistic: 104.8 on 1 and 856 DF,  p-value: < 2.2e-16
```

```
BIC(lm(data$dx_cancer~1)) #Comparing to one
```

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

```
BIC(lm(data$dx_cancer~data$stds_hpv)) #looking at difference
```

```
## [1] -977.596
```

```
#MAJOR CHANGE
```

According to the BIC HPV heavily affects cervical cancer, which was the consensus to start.

Conclusion

After analysis, the data confirmed previous knowledge about cervical cancer, HPV is one of the greatest predictors for cervical cancer. All other factors I measure were almost insignificant compared to HPV. I will say, When it comes to cancer, the safe thing to do is to take all factors into consideration. Any chance to lesson the chances of cancer is a good chance to live longer. Factors such as smoking, and number of sexual partners have very slight affects on cervical cancer, especially compared to HPV. Most people who have cervical cancer are around 33 years old, take hormonal contraceptives. Overall this data set helped me see some of the factors that help contribute toward cervical cancer, and the research for the background information informed me on the stages, and dangers of cervical cancer.

Resources

“Basic Information About Cervical Cancer.” Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, 29 July 2020, www.cdc.gov/cancer/cervical/basic_info/index.htm.

“What Are the Risk Factors for Cervical Cancer?” Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, 7 Aug. 2019, www.cdc.gov/cancer/cervical/basic_info/risk_factors.htm.

“What Can I Do to Reduce My Risk of Cervical Cancer?” Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, 7 Aug. 2019, www.cdc.gov/cancer/cervical/basic_info/prevention.htm.

“HPV Infection.” Mayo Clinic, Mayo Foundation for Medical Education and Research, 30 Aug. 2019, www.mayoclinic.org/diseases-conditions/hpv-infection/symptoms-causes/syc-20351596.

Dillman, Robert K.; Oldham, Robert O., eds. (2009). Principles of cancer biotherapy (5th ed.). Dordrecht: Springer. p. 149. ISBN 9789048122899. Archived from the original on 2015-10-29.

“Basic Information About Cervical Cancer.” Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, 29 July 2020, www.cdc.gov/cancer/cervical/basic_info/index.htm.