Cervical Cancer Risk Prediction

Team 5 - ADS 503

2025-05-31

Introduction

This project is a predictive modeling analysis focused on cervical cancer. The dataset was collected at Hospital Universitario de Caracas in Venezuela and includes patient demographic, lifestyle, and medical history information. The goal is to build models that can predict whether a patient is likely to test positive for cervical cancer based on biopsy outcomes.

Data Importing and Pre-Processing

```
cervical_data <- read_csv("risk_factors_cervical_cancer.csv")
```

```
## Rows: 858 Columns: 36
## -- Column specification ---------------------------------------------------------
## Delimiter: ","
## chr (26): Number of sexual partners, First sexual intercourse, Num of pregna...
## dbl (10): Age, STDs: Number of diagnosis, Dx:Cancer, Dx:CIN, Dx:HPV, Dx, Hin...
##
## 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.
```

```
head(cervical_data, 10)
```

```
## # A tibble: 10 x 36
##      Age ‘Number of sexual partners‘ First sexual interco~1 ‘Num of pregnancies‘
##    <dbl> <chr>                       <chr>                  <chr>
##  1    18 4.0                         15.0                   1.0
##  2    15 1.0                         14.0                   1.0
##  3    34 1.0                         ?                      1.0
##  4    52 5.0                         16.0                   4.0
##  5    46 3.0                         21.0                   4.0
##  6    42 3.0                         23.0                   2.0
##  7    51 3.0                         17.0                   6.0
##  8    26 1.0                         26.0                   3.0
##  9    45 1.0                         20.0                   5.0
## 10    44 3.0                         15.0                   ?
## # i abbreviated name: 1: ‘First sexual intercourse‘
## # i 32 more variables: Smokes <chr>, ‘Smokes (years)‘ <chr>,
## #   ‘Smokes (packs/year)‘ <chr>, ‘Hormonal Contraceptives‘ <chr>,
## #   ‘Hormonal Contraceptives (years)‘ <chr>, IUD <chr>, ‘IUD (years)‘ <chr>,
## #   STDs <chr>, ‘STDs (number)‘ <chr>, ‘STDs:condylomatosis‘ <chr>,
## #   ‘STDs:cervical condylomatosis‘ <chr>, ‘STDs:vaginal condylomatosis‘ <chr>,
## #   ‘STDs:vulvo-perineal condylomatosis‘ <chr>, ‘STDs:syphilis‘ <chr>, ...
```

```
View(head(cervical_data, 10))
```

```
nrow(cervical_data)  # Number of rows (patients)
```

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

```
ncol(cervical_data)  # Number of columns (features)
```

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

```
# cleaning up feature names
cervical_data <- cervical_data %>% clean_names()

# need to manually rename a few features
cervical_data <- cervical_data %>%
  rename(
    stds = st_ds,
    stds_number = st_ds_number,
    stds_condylomatosis = st_ds_condylomatosis,
    stds_cervical_condylomatosis = st_ds_cervical_condylomatosis,
    stds_vaginal_condylomatosis = st_ds_vaginal_condylomatosis,
    stds_vulvo_perineal_condylomatosis = st_ds_vulvo_perineal_condylomatosis,
    stds_syphilis = st_ds_syphilis,
    stds_pelvic_inflammatory_disease = st_ds_pelvic_inflammatory_disease,
    stds_genital_herpes = st_ds_genital_herpes,
    stds_molluscum_contagiosum = st_ds_molluscum_contagiosum,
    stds_aids = st_ds_aids,
    stds_hiv = st_ds_hiv,
    stds_hepatitis_b = st_ds_hepatitis_b,
    stds_hpv = st_ds_hpv,
    stds_number_of_diagnosis = st_ds_number_of_diagnosis,
    stds_time_since_first_diagnosis = st_ds_time_since_first_diagnosis,
    stds_time_since_last_diagnosis = st_ds_time_since_last_diagnosis
  )

colnames(cervical_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"
```

```
# Convert ? to NA
cervical_data[cervical_data == "?"] <- NA
sum(cervical_data == "?", na.rm = TRUE)
```

```
## [1] 0
```

```
# Show unique values for each column
map(cervical_data, ~ unique(.) %>% sort()) %>%
  enframe(name = "column", value = "unique_values") %>%
  print(n = Inf)
```

```
## # A tibble: 36 x 2
##    column                             unique_values
##    <chr>                              <list>
##  1 age                                <dbl [44]>
##  2 number_of_sexual_partners          <chr [12]>
##  3 first_sexual_intercourse           <chr [21]>
##  4 num_of_pregnancies                 <chr [11]>
##  5 smokes                             <chr [2]>
##  6 smokes_years                       <chr [30]>
##  7 smokes_packs_year                  <chr [62]>
##  8 hormonal_contraceptives            <chr [2]>
##  9 hormonal_contraceptives_years      <chr [40]>
## 10 iud                                <chr [2]>
## 11 iud_years                          <chr [26]>
## 12 stds                               <chr [2]>
## 13 stds_number                        <chr [5]>
## 14 stds_condylomatosis                <chr [2]>
## 15 stds_cervical_condylomatosis       <chr [1]>
## 16 stds_vaginal_condylomatosis        <chr [2]>
## 17 stds_vulvo_perineal_condylomatosis <chr [2]>
## 18 stds_syphilis                      <chr [2]>
## 19 stds_pelvic_inflammatory_disease   <chr [2]>
## 20 stds_genital_herpes                <chr [2]>
## 21 stds_molluscum_contagiosum         <chr [2]>
## 22 stds_aids                          <chr [1]>
## 23 stds_hiv                           <chr [2]>
## 24 stds_hepatitis_b                   <chr [2]>
## 25 stds_hpv                           <chr [2]>
## 26 stds_number_of_diagnosis           <dbl [4]>
## 27 stds_time_since_first_diagnosis    <chr [18]>
## 28 stds_time_since_last_diagnosis     <chr [18]>
## 29 dx_cancer                          <dbl [2]>
## 30 dx_cin                             <dbl [2]>
## 31 dx_hpv                             <dbl [2]>
## 32 dx                                 <dbl [2]>
## 33 hinselmann                         <dbl [2]>
## 34 schiller                           <dbl [2]>
## 35 citology                           <dbl [2]>
## 36 biopsy                             <dbl [2]>
```

```
# convert all character columns to numeric
cervical_data <- cervical_data %>%
  mutate(across(where(is.character), ~ as.numeric(.)))

# recategorizing binary indicator variables as categorical (factor) type
# note: All binary variables are coded as 0 = "No" and 1 = "Yes".
factor_vars <- c("smokes", "hormonal_contraceptives", "iud", "stds",
                 "stds_condylomatosis", "stds_cervical_condylomatosis",
                 "stds_vaginal_condylomatosis", "stds_vulvo_perineal_condylomatosis",
                 "stds_syphilis", "stds_pelvic_inflammatory_disease",
                 "stds_genital_herpes", "stds_molluscum_contagiosum",
                 "stds_aids", "stds_hiv", "stds_hepatitis_b", "stds_hpv",
                 "dx_cancer", "dx_cin", "dx_hpv", "dx",
                 "hinselmann", "schiller", "citology", "biopsy")

cervical_data <- cervical_data %>%
  mutate(across(all_of(factor_vars), ~ as.factor(.)))
```

```
head(cervical_data, 10)
```

```
## # A tibble: 10 x 36
##      age number_of_sexual_par~1 first_sexual_interco~2 num_of_pregnancies smokes
##    <dbl>                  <dbl>                  <dbl>              <dbl> <fct>
##  1    18                      4                     15                  1 0
##  2    15                      1                     14                  1 0
##  3    34                      1                     NA                  1 0
##  4    52                      5                     16                  4 1
##  5    46                      3                     21                  4 0
##  6    42                      3                     23                  2 0
##  7    51                      3                     17                  6 1
##  8    26                      1                     26                  3 0
##  9    45                      1                     20                  5 0
## 10    44                      3                     15                 NA 1
## # i abbreviated names: 1: number_of_sexual_partners,
## #   2: first_sexual_intercourse
## # i 31 more variables: smokes_years <dbl>, smokes_packs_year <dbl>,
## #   hormonal_contraceptives <fct>, hormonal_contraceptives_years <dbl>,
## #   iud <fct>, iud_years <dbl>, stds <fct>, stds_number <dbl>,
## #   stds_condylomatosis <fct>, stds_cervical_condylomatosis <fct>,
## #   stds_vaginal_condylomatosis <fct>, ...
```

```
View(head(cervical_data, 10))
```

```r
# view missing data
colSums(is.na(cervical_data))
```

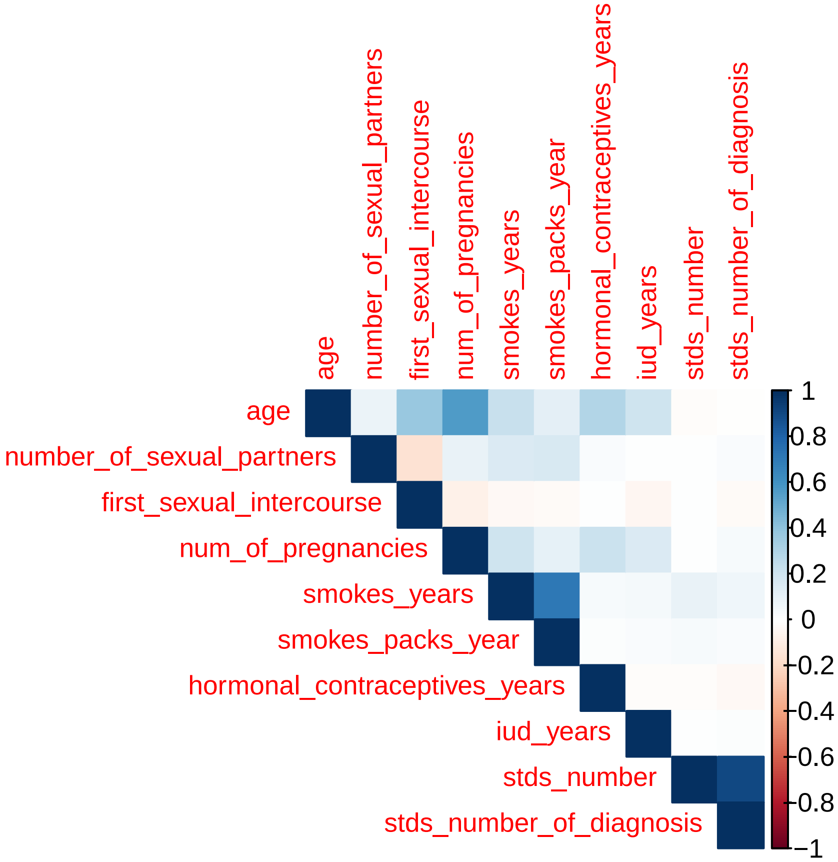
```
##                                age          number_of_sexual_partners 
##                                  0                                 26 
##           first_sexual_intercourse                 num_of_pregnancies 
##                                  7                                 56 
##                             smokes                       smokes_years 
##                                 13                                 13 
##                  smokes_packs_year            hormonal_contraceptives 
##                                 13                                108 
##      hormonal_contraceptives_years                                iud 
##                                108                                117 
##                          iud_years                               stds 
##                                117                                105 
##                        stds_number                stds_condylomatosis 
##                                105                                105 
##       stds_cervical_condylomatosis        stds_vaginal_condylomatosis 
##                                105                                105 
## stds_vulvo_perineal_condylomatosis                      stds_syphilis 
##                                105                                105 
##   stds_pelvic_inflammatory_disease                stds_genital_herpes 
##                                105                                105 
##         stds_molluscum_contagiosum                          stds_aids 
##                                105                                105 
##                           stds_hiv                   stds_hepatitis_b 
##                                105                                105 
##                           stds_hpv           stds_number_of_diagnosis 
##                                105                                  0 
##    stds_time_since_first_diagnosis     stds_time_since_last_diagnosis 
##                                787                                787 
##                          dx_cancer                             dx_cin 
##                                  0                                  0 
##                             dx_hpv                                 dx 
##                                  0                                  0 
##                         hinselmann                           schiller 
##                                  0                                  0 
##                           citology                             biopsy 
##                                  0                                  0
```

```r
# drop sparse columns with >90% missing values (too incomplete for modeling)
cervical_data <- cervical_data %>%
  select(-stds_time_since_first_diagnosis, -stds_time_since_last_diagnosis)
```

```r
# Select numeric columns and remove rows with NAs temporarily
numeric_data <- cervical_data %>%
  select(where(is.numeric)) %>%
  drop_na()

# Compute correlation matrix
cor_matrix <- cor(numeric_data)
```

```r
# Plot correlation heatmap
corrplot(cor_matrix, method = "color", type = "upper", tl.cex = 0.8)
```



```r
# Missing Value Imputation

# separate numeric columns
numeric_vars <- cervical_data %>%
  select(where(is.numeric))

# apply median imputation
preproc <- preProcess(numeric_vars, method = "medianImpute")
numeric_imputed <- predict(preproc, numeric_vars)

# recombine with non-numeric columns (factors)
non_numeric <- cervical_data %>%
  select(where(Negate(is.numeric)))

# Final imputed dataset, and saving separately for models that require no NAs for modeling
cervical_data_imputed <- bind_cols(numeric_imputed, non_numeric)
```

```r
# Create binary numeric version of biopsy (0 = No, 1 = Yes)
cervical_data_imputed$biopsy_numeric <- as.numeric(cervical_data_imputed$biopsy) - 1

# Select numeric predictors (exclude biopsy_numeric itself)
numeric_vars <- cervical_data_imputed %>%
  select(where(is.numeric)) %>%
  select(-biopsy_numeric)

# Compute point-biserial correlation between each predictor and biopsy_numeric
cor_results <- sapply(numeric_vars, function(x) {
  cor.test(cervical_data_imputed$biopsy_numeric, x)$estimate
})

# Display sorted correlations from strongest to weakest
sort(cor_results, decreasing = TRUE)
```

```
##                   stds_number.cor      stds_number_of_diagnosis.cor
##                      0.1031527184                      0.0974489209
## hormonal_contraceptives_years.cor                  smokes_years.cor
##                      0.0941636403                      0.0612042289
##                           age.cor            num_of_pregnancies.cor
##                      0.0559555151                      0.0402150719
##                     iud_years.cor             smokes_packs_year.cor
##                      0.0381761371                      0.0244868673
##      first_sexual_intercourse.cor     number_of_sexual_partners.cor
##                      0.0072587257                     -0.0004082348
```

The correlation is all very weak, all less than 0.1.

Exploratory Data Analysis

```
summary(cervical_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
##                  NA’s   :26                NA’s   :7
##  num_of_pregnancies  smokes     smokes_years   smokes_packs_year
##  Min.   : 0.000     0   :722   Min.   : 0.00   Min.   : 0.0000
##  1st Qu.: 1.000     1   :123   1st Qu.: 0.00   1st Qu.: 0.0000
##  Median : 2.000     NA’s: 13   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
##  NA’s   :56                    NA’s   :13      NA’s   :13
##  hormonal_contraceptives hormonal_contraceptives_years   iud
##  0   :269                Min.   : 0.000                0   :658
##  1   :481                1st Qu.: 0.000                1   : 83
##  NA’s:108                Median : 0.500                NA’s:117
##                          Mean   : 2.256
##                          3rd Qu.: 3.000
##                           Max.   :30.000
##                           NA's   :108
##    iud_years         stds      stds_number     stds_condylomatosis
##  Min.   : 0.0000   0   :674   Min.   :0.0000   0   :709
##  1st Qu.: 0.0000   1   : 79   1st Qu.:0.0000   1   : 44
##  Median : 0.0000   NA's:105   Median :0.0000   NA's:105
##  Mean   : 0.5148              Mean   :0.1766
##  3rd Qu.: 0.0000              3rd Qu.:0.0000
##  Max.   :19.0000              Max.   :4.0000
##  NA's   :117                  NA's   :105
##  stds_cervical_condylomatosis stds_vaginal_condylomatosis
##  0   :753                     0   :749
##  NA's:105                     1   :  4
##                               NA's:105
##
##
##
##
##  stds_vulvo_perineal_condylomatosis stds_syphilis
##  0   :710                           0   :735
##  1   : 43                           1   : 18
##  NA's:105                           NA's:105
##
##
##
##
##  stds_pelvic_inflammatory_disease stds_genital_herpes
##  0   :752                         0   :752
##  1   :  1                         1   :  1
##  NA's:105                         NA's:105
##
##
##
##
##  stds_molluscum_contagiosum stds_aids  stds_hiv   stds_hepatitis_b stds_hpv
##  0   :752                   0   :753   0   :735   0   :752         0   :751
##  1   :  1                   NA's:105   1   : 18   1   :  1         1   :  2
##  NA's:105                              NA's:105   NA's:105         NA's:105
##
##
##
##
##  stds_number_of_diagnosis dx_cancer dx_cin  dx_hpv  dx      hinselmann schiller
##  Min.   :0.00000          0:840     0:849   0:840   0:834   0:823      0:784
##  1st Qu.:0.00000          1: 18     1:  9   1: 18   1: 24   1: 35      1: 74
##  Median :0.00000
##  Mean   :0.08741
##  3rd Qu.:0.00000
##  Max.   :3.00000
##
##  citology biopsy
##  0:814    0:803
##  1: 44    1: 55
##
```

```
##
##
##
##
```

```r
# drop features with only one unique value and NA: no predictive power
cervical_data <- cervical_data %>%
  select(-stds_cervical_condylomatosis, -stds_aids)
```

We found that stds_cervical_condylomatosis and stds_aids each contained only one unique non-missing value (all "0", or "No") and had a high proportion of missing values (NA). This indicates that they provide no meaningful variation for modeling and would add unnecessary sparsity to the data. While it's possible that the missingness is related to the value itself (e.g., respondents choosing not to disclose due to sensitivity), the lack of variation in the observed data makes it impossible to model these features reliably. This aligns with guidance that missingness dependent on the unobserved value itself (i.e., Not Missing At Random) presents a particularly difficult modeling scenario (Kuhn & Johnson, 2013). **Citing WEEK 2 Discussion post reading, include in final paper!

```r
# we want to do more discovery in selecting the target variable
table(cervical_data$biopsy)
```

```
##
##   0   1
## 803  55
```

```r
table(cervical_data$dx_cancer)
```

```
##
##   0   1
## 840  18
```

This table tells us that biopsy positives are more common than positive cancer diagnoses. Only 18 patients were diagnosed with cancer, which is very small for a classification target.

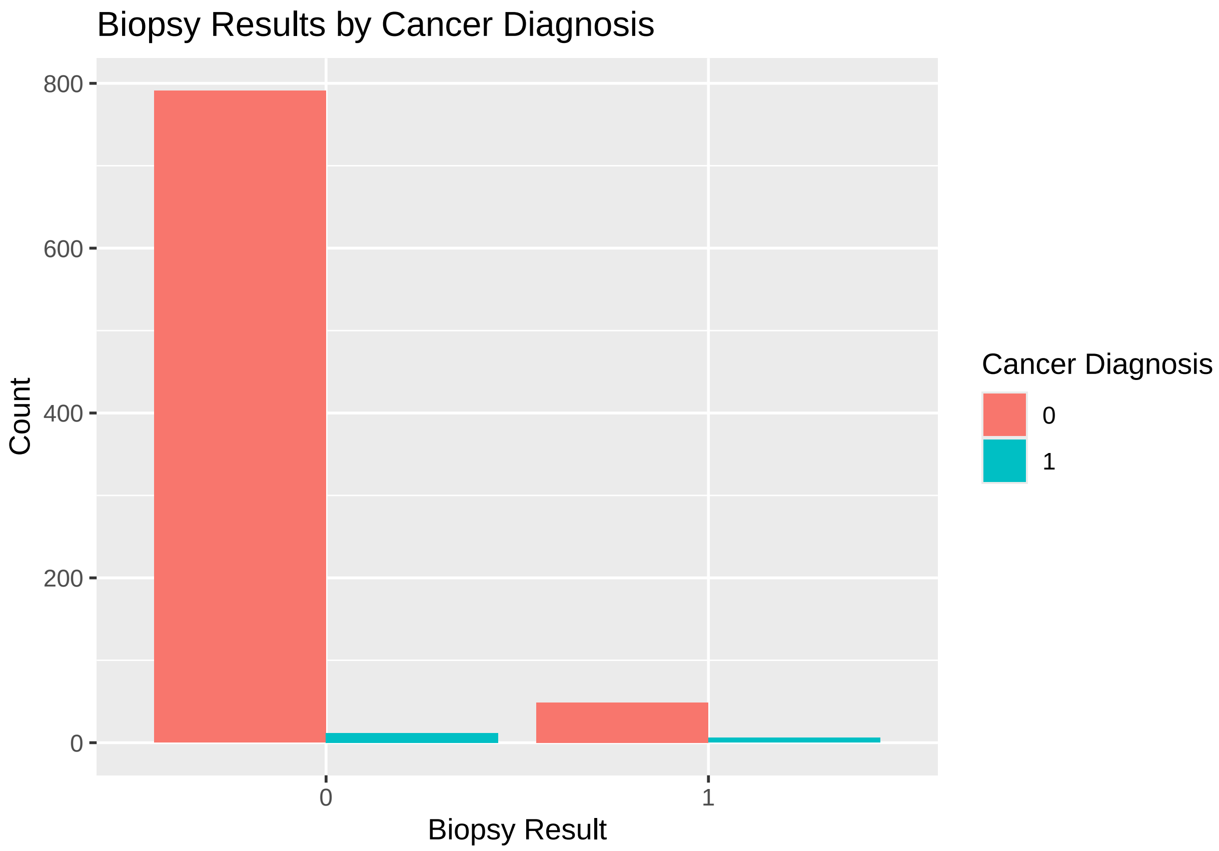
```r
tab <- table(cervical_data$biopsy, cervical_data$dx_cancer)

dimnames(tab) <- list(
  Biopsy = c("Negative", "Positive"),
  CancerDiagnosis = c("No Cancer", "Cancer")
)
addmargins(tab)
```

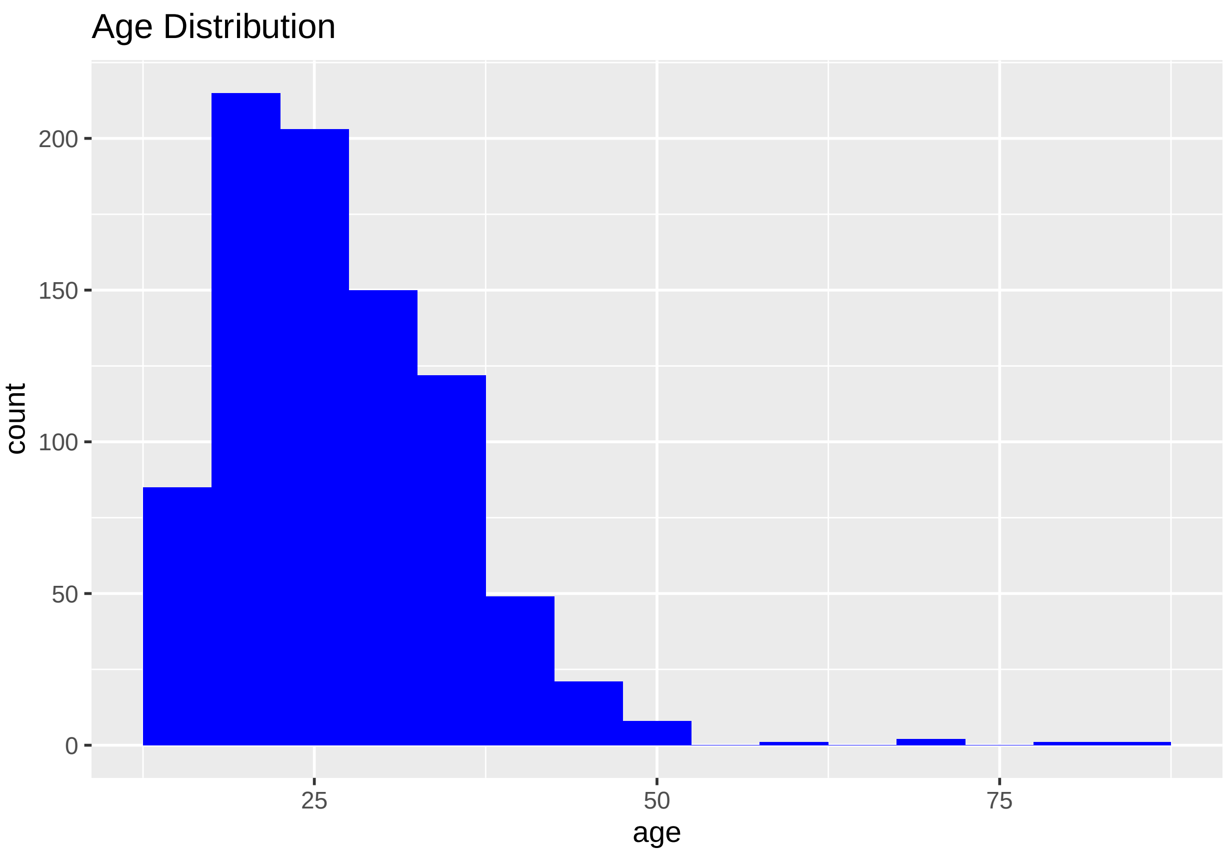
```
##             CancerDiagnosis
## Biopsy       No Cancer Cancer Sum
##   Negative         791     12 803
##   Positive          49      6  55
##   Sum              840     18 858
```

To note from this table: Not all positive biopsies were diagnosed as cancer. 49 patients had a positive biopsy, but no confirmed cancer diagnosis (could be precancerous). Some patients were diagnosed with cancer despite a negative biopsy (could be a preexisting diagnosis).

```
ggplot(cervical_data, aes(x = biopsy, fill = dx_cancer)) +
  geom_bar(position = "dodge") +
  labs(title = "Biopsy Results by Cancer Diagnosis",
       x = "Biopsy Result", y = "Count", fill = "Cancer Diagnosis")
```

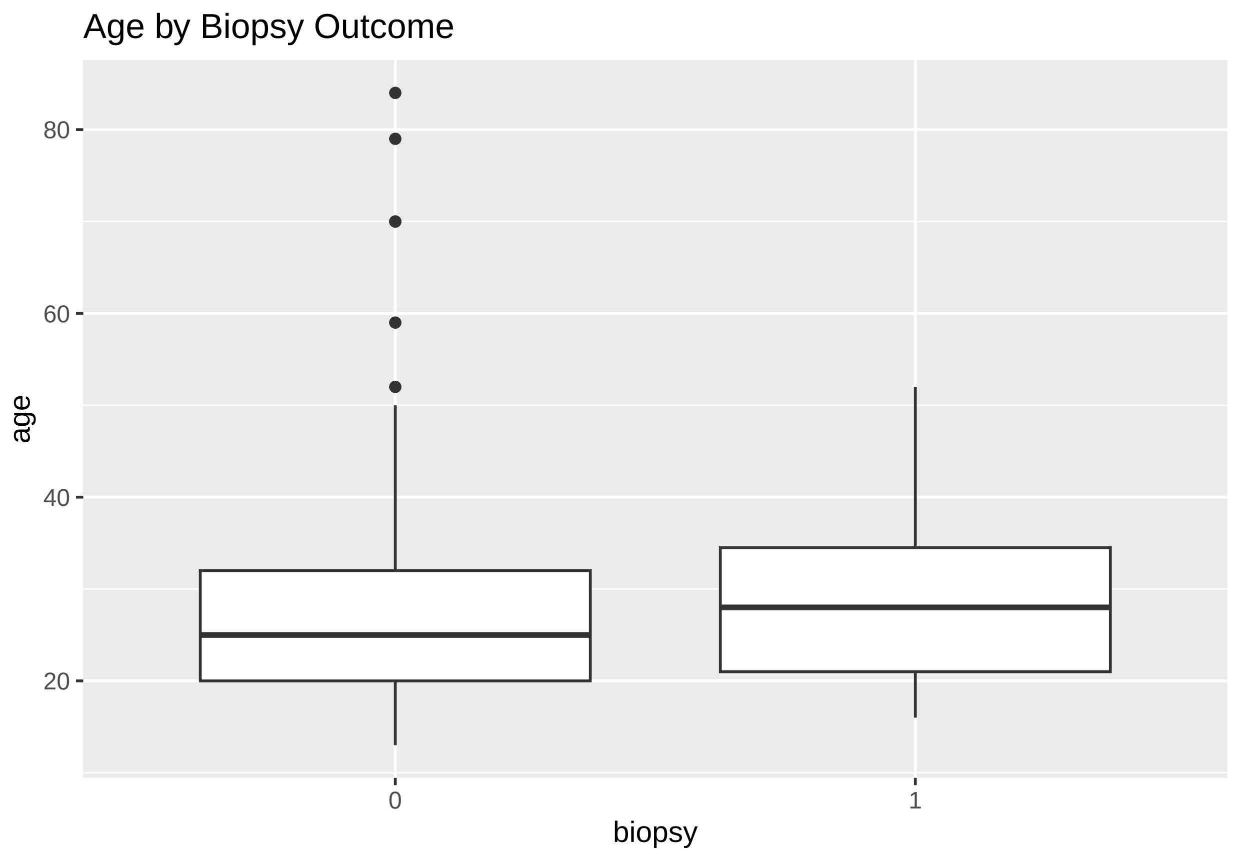


```
ggplot(cervical_data, aes(x = age)) +
  geom_histogram(binwidth = 5, fill = "blue") +
  labs(title = "Age Distribution")
```



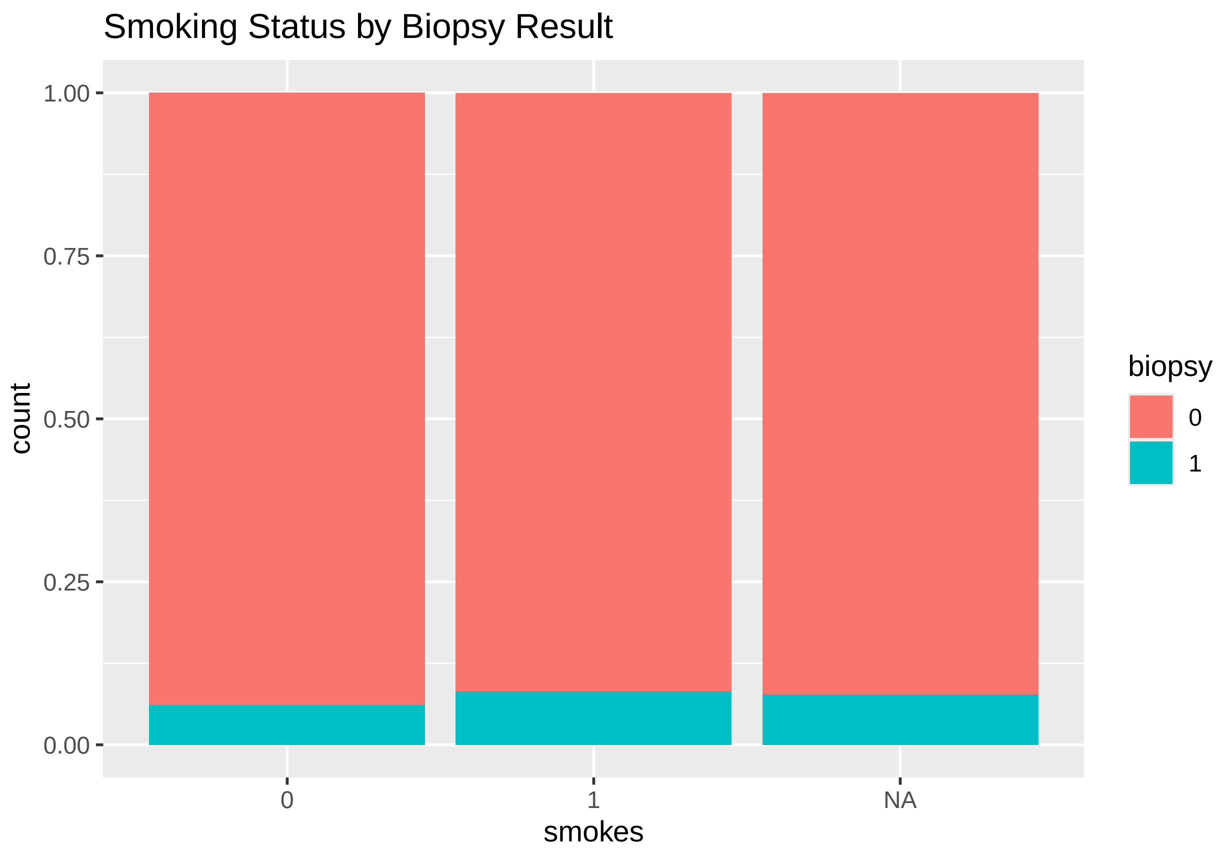
The age distribution of patients is right-skewed, with the majority of patients between 20 and 40 years old. A small number of patients are over 50, and very few are over 70. This suggests that the dataset largely reflects a younger population, which aligns with the typical age range for cervical cancer screening. However, outliers in the older range could be important to monitor for elevated risk patterns.

```r
ggplot(cervical_data, aes(x = biopsy, y = age)) +
  geom_boxplot() +
  labs(title = "Age by Biopsy Outcome")
```



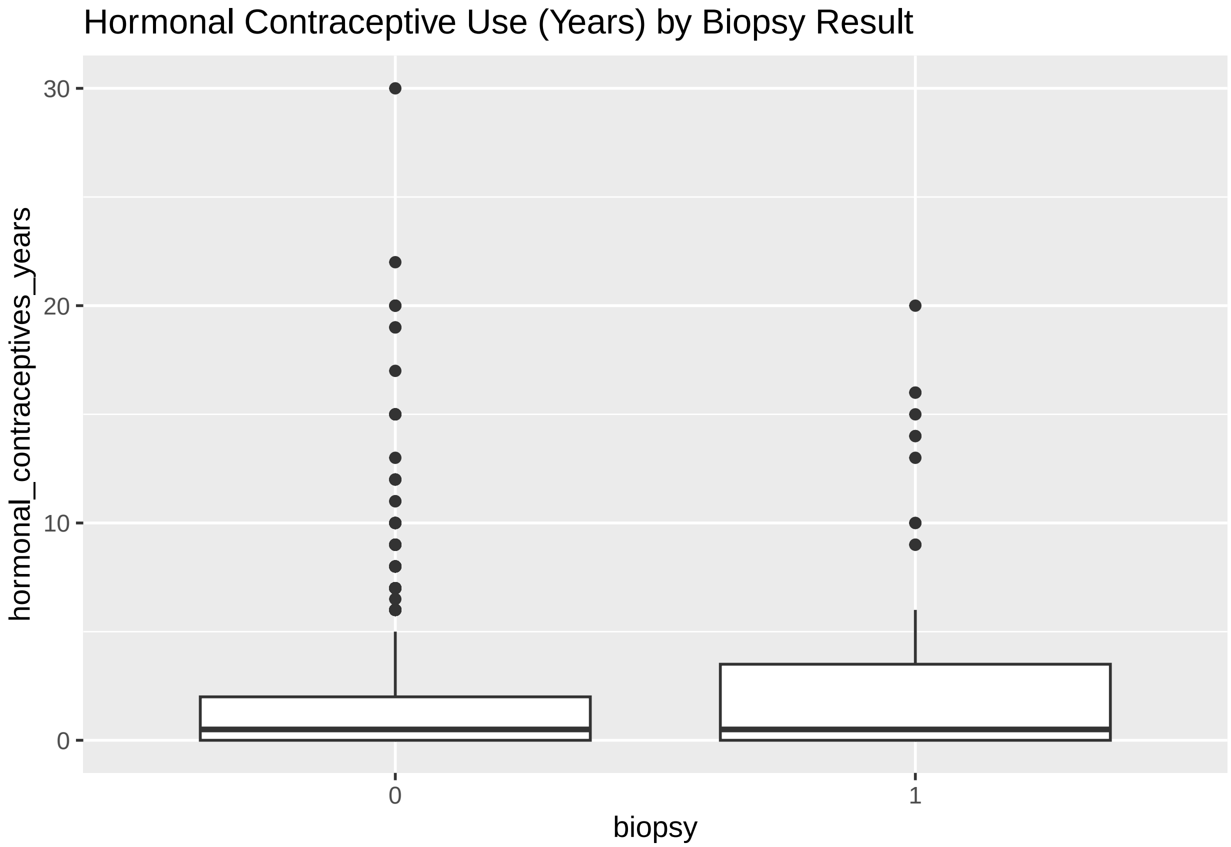
The median age for patients with a positive biopsy result appears slightly higher than for those with a negative result. While there is considerable overlap in the distributions, the boxplot suggests that older patients may be more likely to test positive for cervical cancer. A few older individuals with negative biopsy results appear as outliers, but the positive group shows a more concentrated distribution between ages 25 and 45.

```
ggplot(cervical_data, aes(x = smokes, fill = biopsy)) +
  geom_bar(position = "fill") +
  labs(title = "Smoking Status by Biopsy Result")
```



This bar plot shows the proportion of biopsy outcomes (0 = Negative, 1 = Positive) across smoking status groups (0 = non-smoker, 1= smoker, NA = missing value). The distribution of biopsy results appears very similar across all three smoking categories. This suggests that smoking status does not show a strong relationship with biopsy outcome in this dataset.

```r
# view the distribution of years of hormonal contraceptive use across biopsy (using the imputed dataset
ggplot(cervical_data_imputed, aes(x = biopsy, y = hormonal_contraceptives_years)) +
  geom_boxplot() +
  labs(title = "Hormonal Contraceptive Use (Years) by Biopsy Result")
```



Although the medians for hormonal contraceptive use are similar between biopsy outcome groups, the distribution for the positive biopsy group is more spread out and contains more patients with slightly longer years of use. This could indicate that patients with positive biopsy results may have a few more years of contraceptive use (although correlation was weak).