Cervical Cancer Interpolation and Prediction

Wai (Antina) Lee
Joey Haymaker
Maria Diaz Ortiz

12/13/2017

Contents

Executive Summary **2**

Description of the Problem **2**

Description of the Data **3**
Predictor variables . . . 3
Response Variable (y) . . . 4

Data Cleaning and Preparation (EDA) **4**
Dependent variable `cancerPred` . . . 4
Smoking . . . 6
Age . . . 12
Sexual activity . . . 13
HPV . . . 16
Multicollinearity across variables . . . 17
Final dataset, Testing & Training Data Split . . . 23

Modelling **24**
Logistic Regression with LASSO . . . 24
Random Forest . . . 24
Comparison of Methods . . . 25
Validity of Results and Future Improvement . . . 27

Conclusion **28**

Appendix **28**
Original variable descriptions . . . 28

Executive Summary

Cervical cancer is an important clinical problem both in the US and worldwide, and is associated with human papilloma virus (HPV). In order to decrease its incidence, it is important to inderstand what factors affect one's likelihood to develop cervical cancer. Our goal was to identify factors that significantly impact the risk of cancer by building a model that predicts whether a patient is "high risk" or "low risk" for developing the disease. We explored a kaggle dataset of 858 patients and 26 variables, which included known risk factors for cervical cancer. Rows of missing values, highly correlated variables, and sparse variables were eliminated and relevant variables were log(transfored). The response variable `cancerPred` was developed as a sum of 4 diagnostic variables (ranging from 0-4). To perform prediction, the response variable `cancerPred` was further subdivided into "low risk" or "0" (for an initial `cancerPred` score of 0) and "high risk" or "1" (for an initial`cancerPred` score of 1-4). The resulting dataset was split 70-30 into training and testing datasets. LASSO with 10-fold cross-validation was performed to identify non-zero coefficients, which were used to perform logistic regression with backwards selection until all coefficients were significant at the alpha = 0.05 level. A second classifier was built using randomForest with an mtry of 4 on the training data. Both fits were used to predict $P(\hat{Y} = 1)$ for the test data, then $P(\hat{Y} = 1) > 1/4$ was used as the threshold rule to determine class assignments. Sensitivity, specificity, false nagative rate, false positive rate and unweighted misclassification errors (MCE) were calculated for the testing data to compared model performance for the selected rule. ROC curves and AUC were also calculated to compare performance across all rules. Random forest outperformed logistic regression in classifying our training data (with an AUC of near 1); however, both models had similar performance in terms of AUC for the test data (both around ~0.58). Although logistic regression outperformed random forest in terms of MCE for the specified rule, random forest had a higher sensitivity ($P(\hat{Y} = 1|Y = 1)$) and lower false negative rate ($P(\hat{Y} = 0|Y = 1)$), which are two desirable qualities when classifying potential cancer patients. Future directions for improvement include significantly increasing our sample size (particuarly in the proportion of "high risk" patients) and exploring the effect of vaccination against HPV.

Description of the Problem

The incidence of cervical cancer in the US has decreased singnificanty over the past 40 years (from being the #1 cause of death for women in the US to being 14th in frequency), mosly due to early detection from pap smear screening. However, it still remains a large source of morbidity and mortality both in the US and globally. In 2014, there were over 12,500 new cervical cancer diagnoses and over 4,000 cervical cancer-associated deaths in the US alone. On a global scale, cervical cancer is still the 3rd most common neoplasia and the #2 cause of cancer-related deaths among women. Cervical cancer is almost exclusively caused by the human papilloma virus (HPV), a common sexually transmitted disease (STD). There are multiple subtypes of HPV, some of which are associated with warts (condyloma/condylomatosis) and others which are notorious for causing cervical cancer. Although the US Food and Drug Administration (FDA) has approved two vaccines for HPV, these vaccines do not protect against all cancer-causing HPV strains. Thus, in order to further decrease the incidence of this malignancy, it is important to understand what behavioral and demographic factors increase the likelihood of acquiring HPV and developing cervical cancer. Some factors that are traditionally associated with an increased risk include an earlier age for 1st sexual encounter and a larger number of sexual partners (since both increase likelihood of being exposed to HPV), as well as being of older age (since cancer takes a couple of years to develop after exposure and younger women's bodies tend to heal from HPV without progressing to CIN or cancer). Thus, the goal of our analysis is to develop a model that can estimate the risk of developing cervical cancer using a patient derived datase. We hope to validate some of the previous findings in the literature as well as to expand upon this knowledge using the results from our approach.

Description of the Data

The dataset was obtained from kaggle and was collected by administering questionnaires to 858 patients at 'Hospital Universitario de Caracas' in Caracas, Venezuela. It includes demographic information, habits, and historic medical records of the patients in question. Missing values correspond to patients who failed to answer a question due to privacy concerns. After EDA, cleaning and transformations, a total of 12 variables were used to predict the response variable. A description of all the variables in the original dataset are included in the appendix.

Predictor variables

Demographic

- `logAge` (numeric) - Corresponds to the age at which the patient filled the questionnaire, log transformed for normality
- `Smoker.status` (factor) - Categorical value that corresponds to whether someone is a non-smoker (0), light-medium smoker (1), or heavy smoker (2).
  - This variable was generated by using the following transformation to define pack years (a metric commonly used in the medical field to predict one's risk of other diseases, including lung cancer): $Packyears = Smokes(yes = 1, no = 0) \times Smokes(years) \times Smokes(packs/year)$. Then, pack years was log transformed ($log(packyears + 1)$) because, when excluding non-smokers ($log(packyears + 1) = 0$), the distribution for number of $log(packyears + 1)$ looks more normal. We calculated the mean for this group ($mean(log(packyears + 1)) = 2.346$) and 0-2 were assigned as follows:
    - $Non - smoker(0)$: 0 log(packyears+1)
    - $Light - Mediumsmoker(1)$: Between 0 and 2.346 log(packyears+1)
    - $Heavysmoker(2)$: > 2.346 log(packyears+1)

Contraception

- `log.Years.HCP` (numeric)** - Generated by multiplying $Years.HCP = Hormonal.Contraceptives(yes = 1, no = 0) \times Hormonal.Contraceptive.Years$ then taking the log transofrmation: $log.Years.HCP = log(Years.HCP + 1)$
- `log.Years.IUD` (numeric)** - Generated by multiplying $Years.IUD = IUD(yes = 1, no = 0) \times IUD.Years$ then taking the log transofrmation: $log.Years.IUD = log(Years.IUD+1)$ ** The variables `log.Years.HCP` and `log.Years.IUC` were used instead of `Years.HCP` and `Years.IUD` respectively because, when you eliminate patients without HCP and IUD's, the log distributions are closer to being normally distributed (see EDA below).

Sexual activity

- `Number.of.sexual.partners` (numeric) - Numer of sexual partners
- `First.sexual.intercourse` (numeric) - Age at first sexual intercourse
- `Num.of.pregnancies` (numeric) - Number of pregnancies

Sexual health history

- `STDs.condylomatosis` (factor) - Categorical variable representing the presence ($STDs.condylomatosis = 1$) or absence ($STDs.condylomatosis = 0$) of genital warts, which is a sexually transmitted disease caused by HPV strains that are not cancer-associated.
- `STDs.syphilis` (factor) - Categorical variable representing the presence ($STDs.syphilis = 1$) or absence ($STDs.syphilis = 0$) of syphilis. This STD is caused by a bacteria rather than a virus, and is relativey rare in comparison to HPV.

Medical history

- `Dx.Cancer` (factor) - Categorical variable corresponding to whether or not a person carries a previous cervical cancer diagnosis (1) or not (0)
- `Dx.CIN` (factor) - Categorical variable corresponding to whether or not a person carries a previous diagnosis of cervical intraepithelial neoplasia or CIN (1) or not (0). CIN is when a biopsy of the cervix is examined under the microscope and reveals malignant or cancer-like cells that have NOT invaded outside of a well defined area. This is considered a form or "pre-cancer" and is thought to progress to cancer if left untreated
- `Dx.HPV` (factor) - Categorical variable corresponding to whether or not a person has been previously diagnosed with HPV (1) or not (0). Although HPV is the #1 risk factor for cancer, there is not a 1-to-1 relationship, since many women's bodies are able to heal themselves and become HPV negative over time (especially young people)

Response Variable (y)

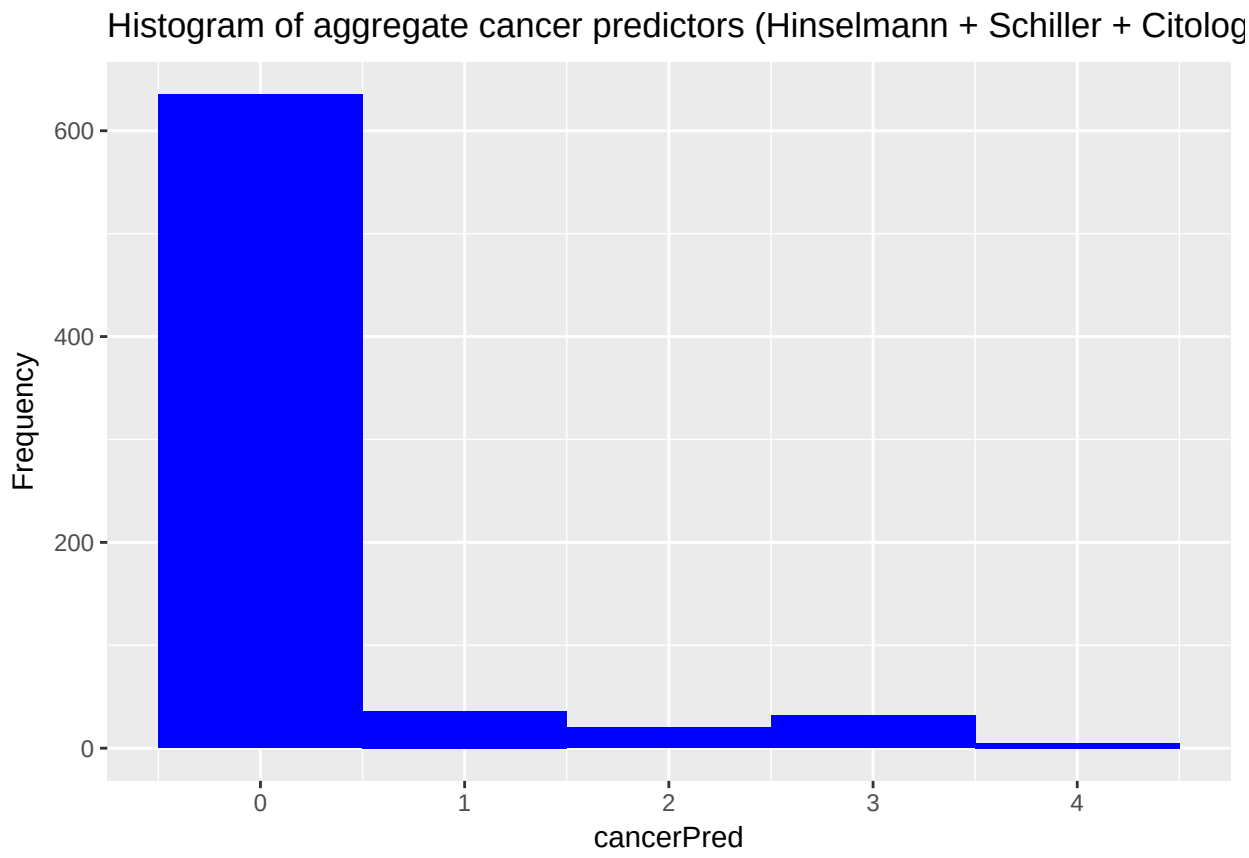
- `cancerPred` (factor) - Categorical variable generated as the sum of 4 individual variables ($Schiller + Hinselmann + Cytology + Biopsy$) which is meant to capture a person's risk of having cervical cancer. More detailed descriptions of what the variables mean are included below:
  - *Schiller*: The Schiller test is a screening test in which iodine solution is added to the cervix to better visualize any abnormal cells that may be suspicious for cancer. A positive test ($Schiller = 1$) increases your risk of cervical cancer and might require further diagnostic exams. On the other hand, a negative test ($Schiller = 0$) makes having cervical cancer less likely.
  - *Hinselmann*: The Hinselmann tets is more commonly used suring screening and involves adding an acetic acid or vinegar solution to the cervix to visualize abnormal cells. An abnormal test ($Hinselmann = 1$) increases your risk of cervical cancer and usually warrants either cytology or a biopsy, whereas a negative test ($Hinselmann = 0$) decreases your likelihood of having cervical cancer.
  - *Cytology*: This is when cells are taken from the cervix using a brush, then they aresuspended in a liquid solution and visualized under a microscope.

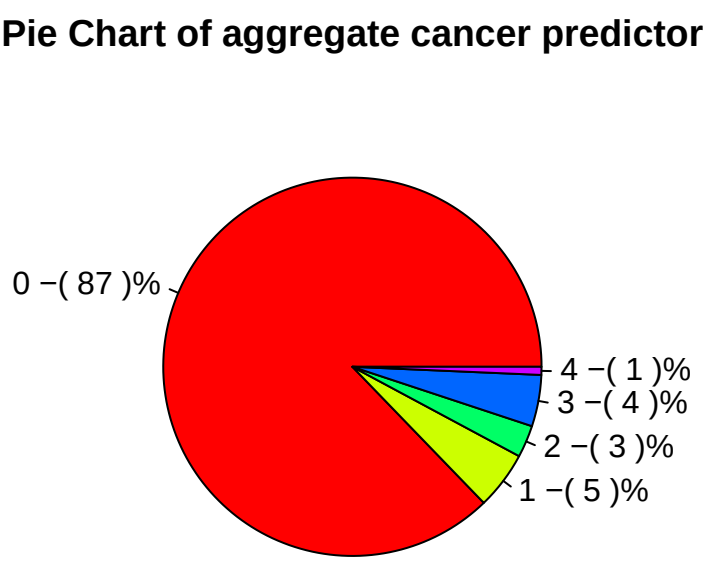
Data Cleaning and Preparation (EDA)

While there were many good variables that theoretically would have been good to include in our model (such as `STDs.HIV`, `STDs.Hepatitis.B`, `STDs.HPV`), many were almost purely comprised of 0 values, leaving little to be inferred from them. Consequently, they were removed from the dataset. There were a large number of '?' entries, many of which were also eliminated from the dataset. This left 728 out of the original 858 observations in the dataset.

Dependent variable `cancerPred`

The response variable is created from aggregating the last four features `Hinselmann`, `Schiller`, `Citology`, and `Biopsy`, as they represent medical exams that determine the likelyhood of that person having cervical cancer.

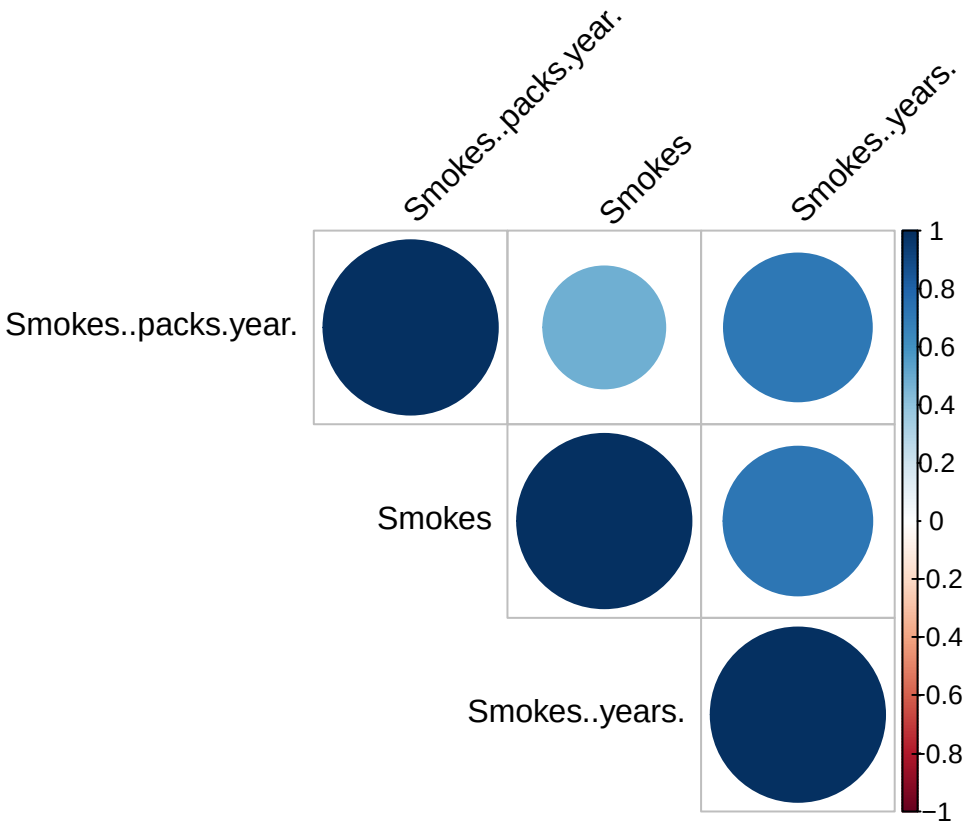

Histogram of aggregate cancer predictors (Hinselmann + Schiller + Citolog
Frequency
600
400
200
0
0
1
2
3
4
cancerPred



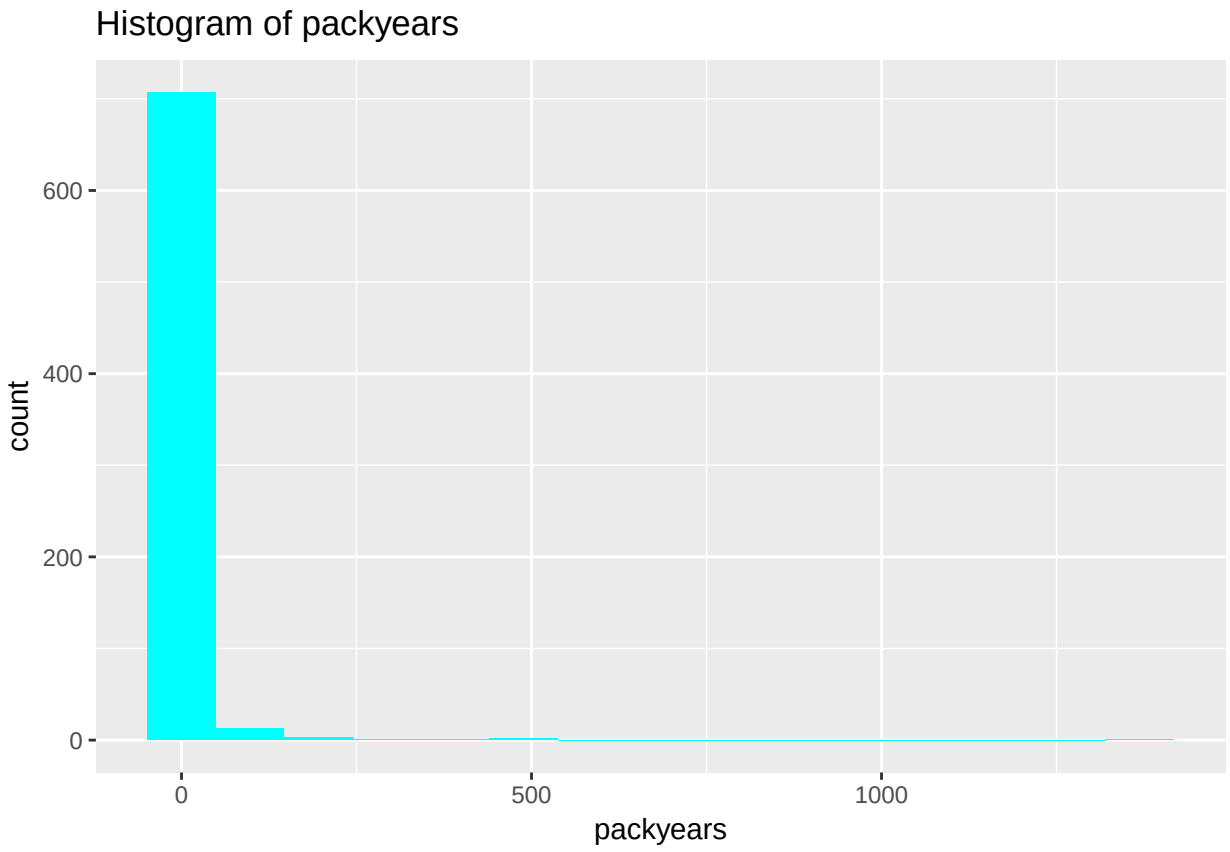
There are a great deal more values that equal 0 than any other value.

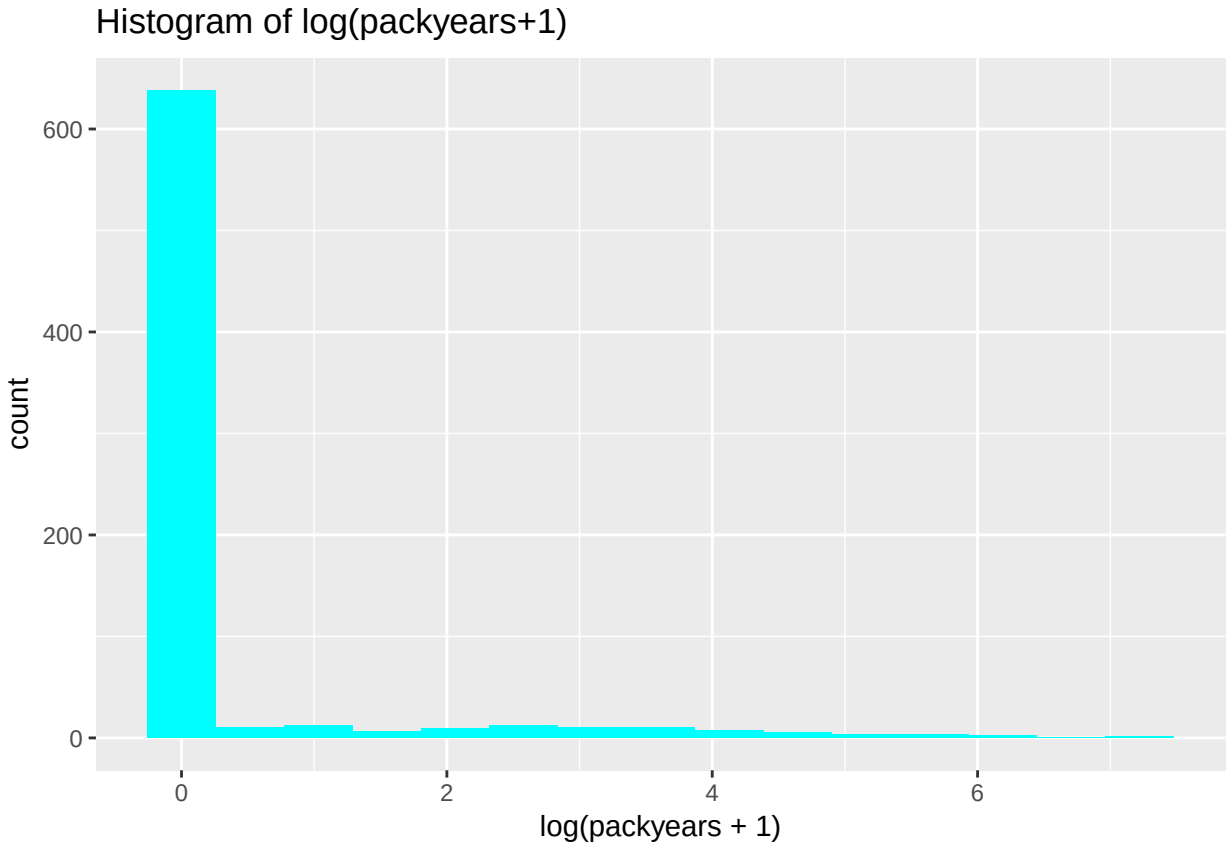
Smoking

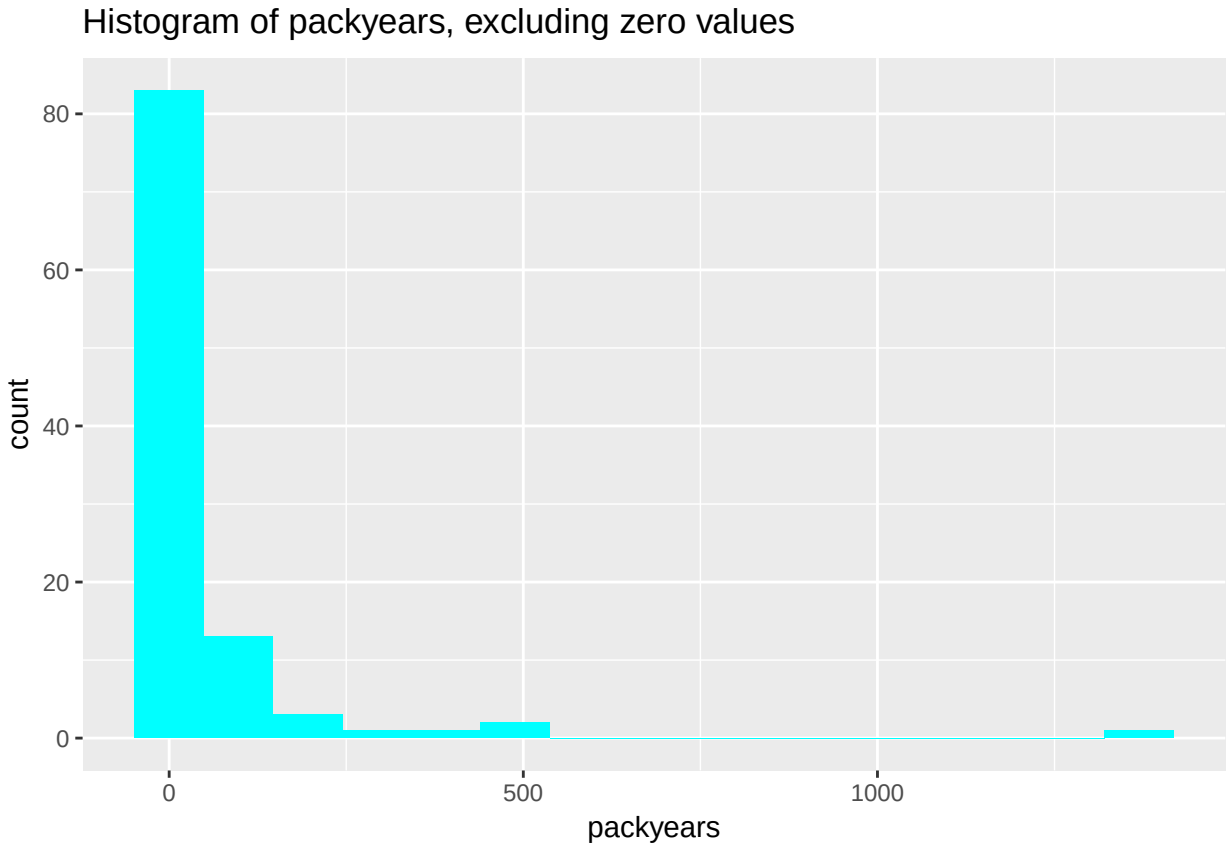
Captured in: `Smokes`, `Smokes..years.`, `Smokes..packs.year.`

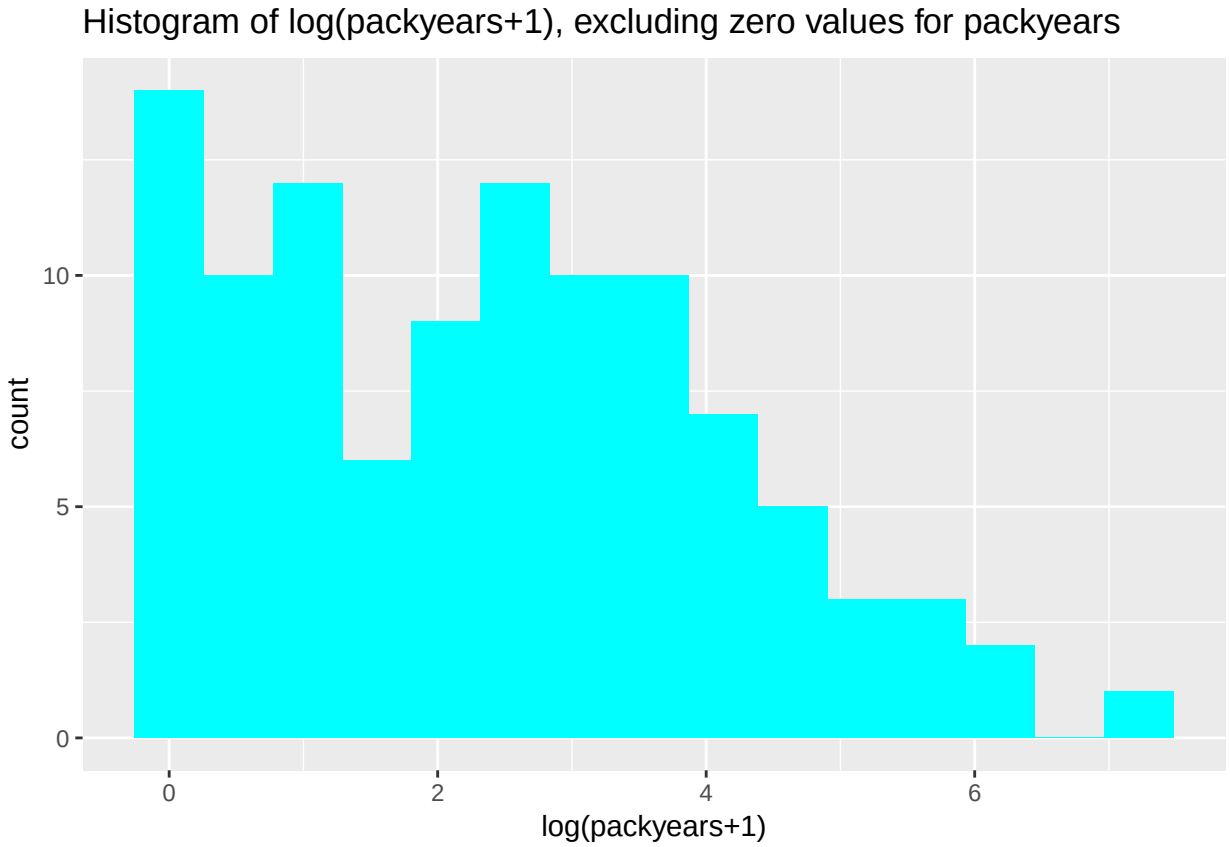


All these variables are highly correlated and could be combined.


Histogram of packyears
count
600
400
200
0
0
500
1000
packyears


Histogram of log(packyears+1)
count
600
400
200
0
0
2
4
6
log(packyears + 1)

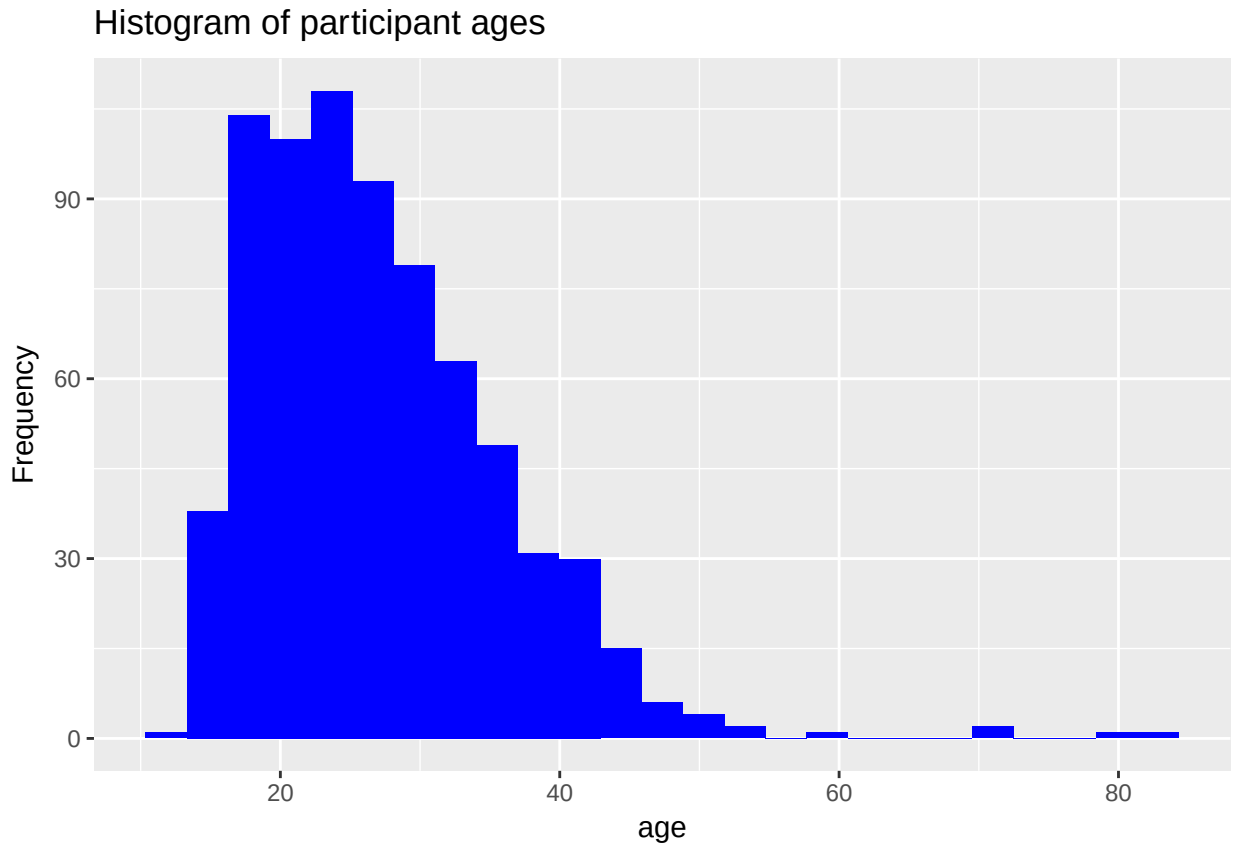

Histogram of packyears, excluding zero values
count
80
60
40
20
0
0
500
1000
packyears

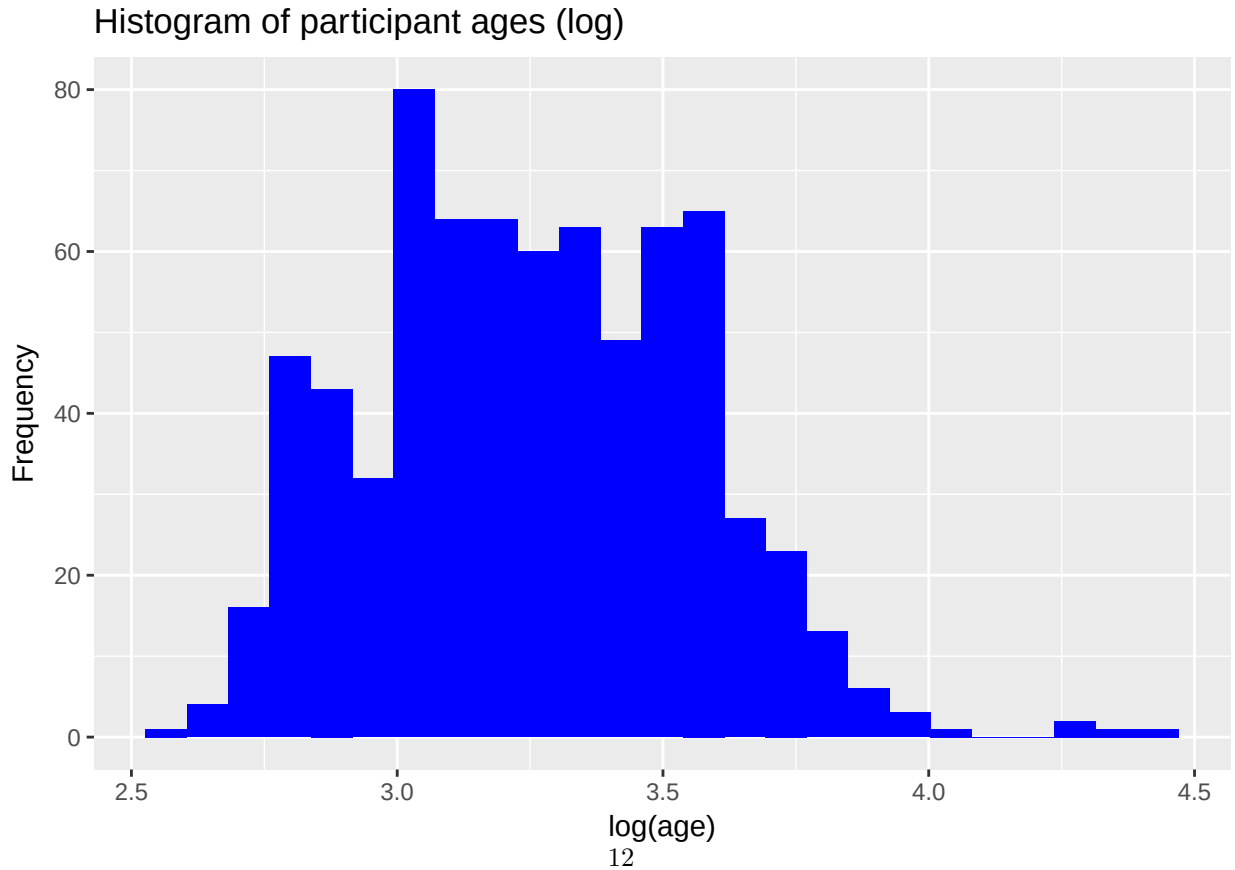


To address the interaction between the 3 smoking variables, a categorical variable `Smoker.status` was created. This was done first by a transformation to define pack years (a metric commonly used in the medical field to predict one's risk of other diseases, including lung cancer): $Packyears = Smokes(yes = 1, no = 0) \times Smokes(years) \times Smokes(packs/year)$. Then, pack years was log transformed ($log(packyears + 1)$) because, when excluding non-smokers ($log(packyears + 1) = 0$), the distribution for number of $log(packyears + 1)$ looks more normal (as seen above). Finally, the mean was calculated for the group ($mean(log(packyears + 1)) = 2.346$) and 0-2 were assigned as follows: $+ Non-smoker(0)$: 0 log(packyears+1) $+ Light-Mediumsmoker(1)$: Between 0 and 2.346 log(packyears+1) $+ Heavysmoker(2)$: > 2.346 log(packyears+1)

The original smoking variables (`Smokes`, `Smokes..years.`, `Smokes..packs.year.`, `packyears`) were then removed fromt the dataset.

Age


Histogram of participant ages
Frequency
90
60
30
0
20
40
60
80
age

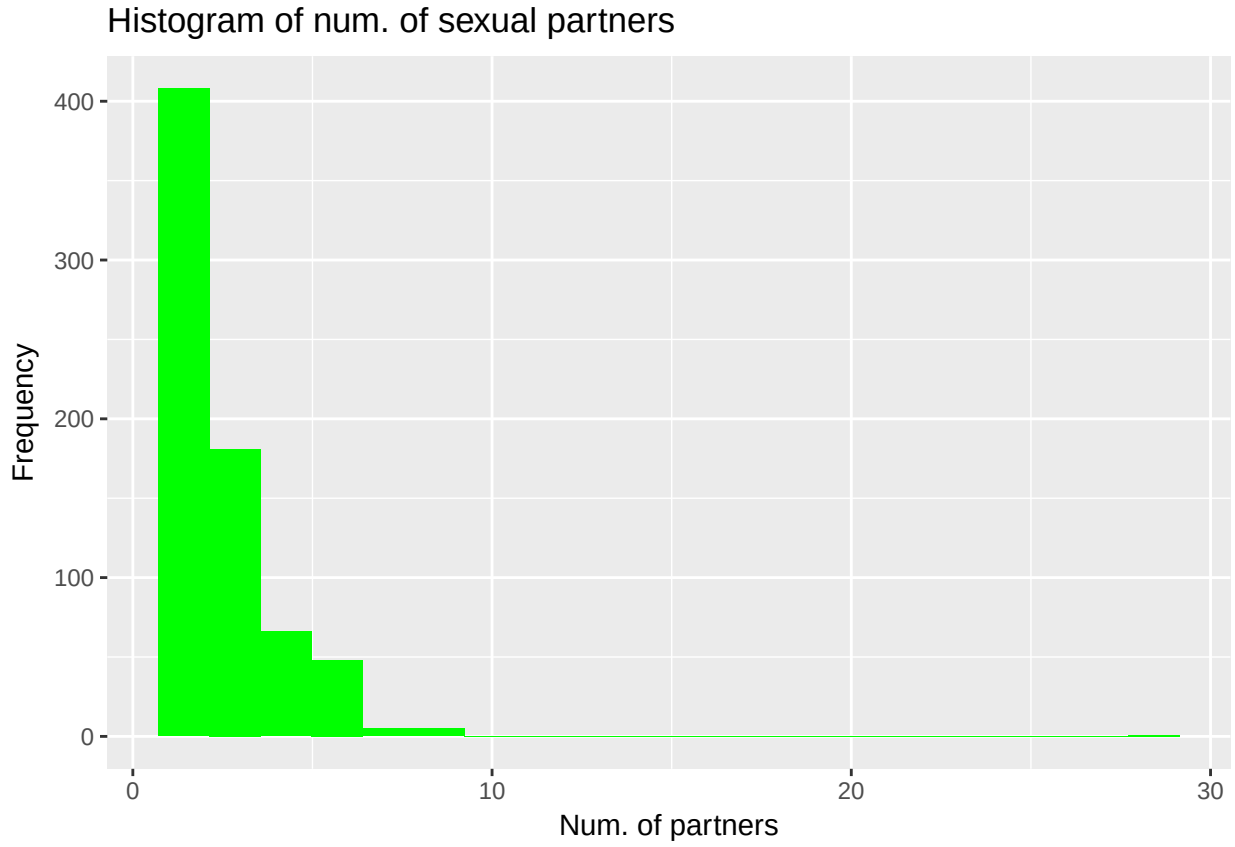

Histogram of participant ages (log)
Frequency
80
60
40
20
0
2.5
3.0
3.5
4.0
4.5
log(age)

We can see the respondents' ages are skewed, with the majority of participants falling in the 20-40 age range. There is a conspicuos gap in representation, specifically in teh 60-70 age range, as well as 70-80 age range.

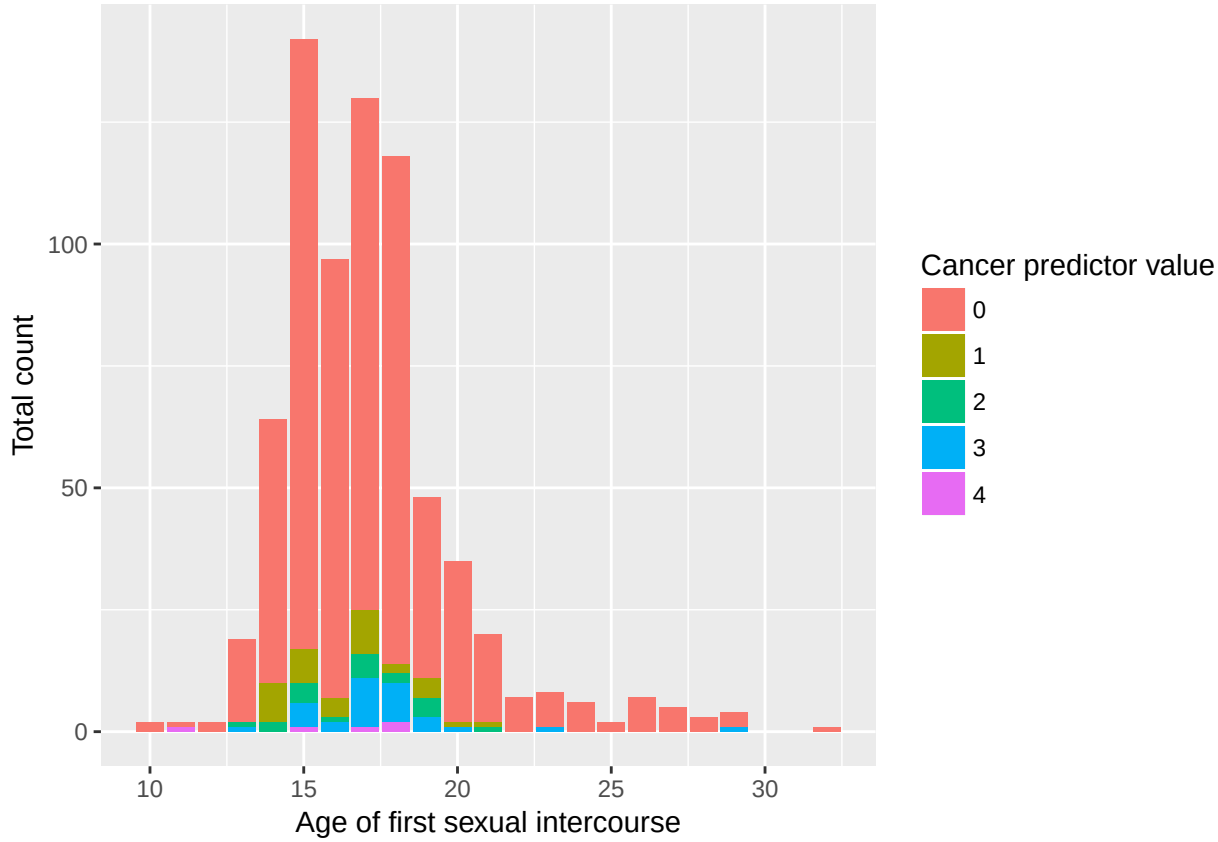
Taking the natural log of the age results in a more normal distribution.

Sexual activity

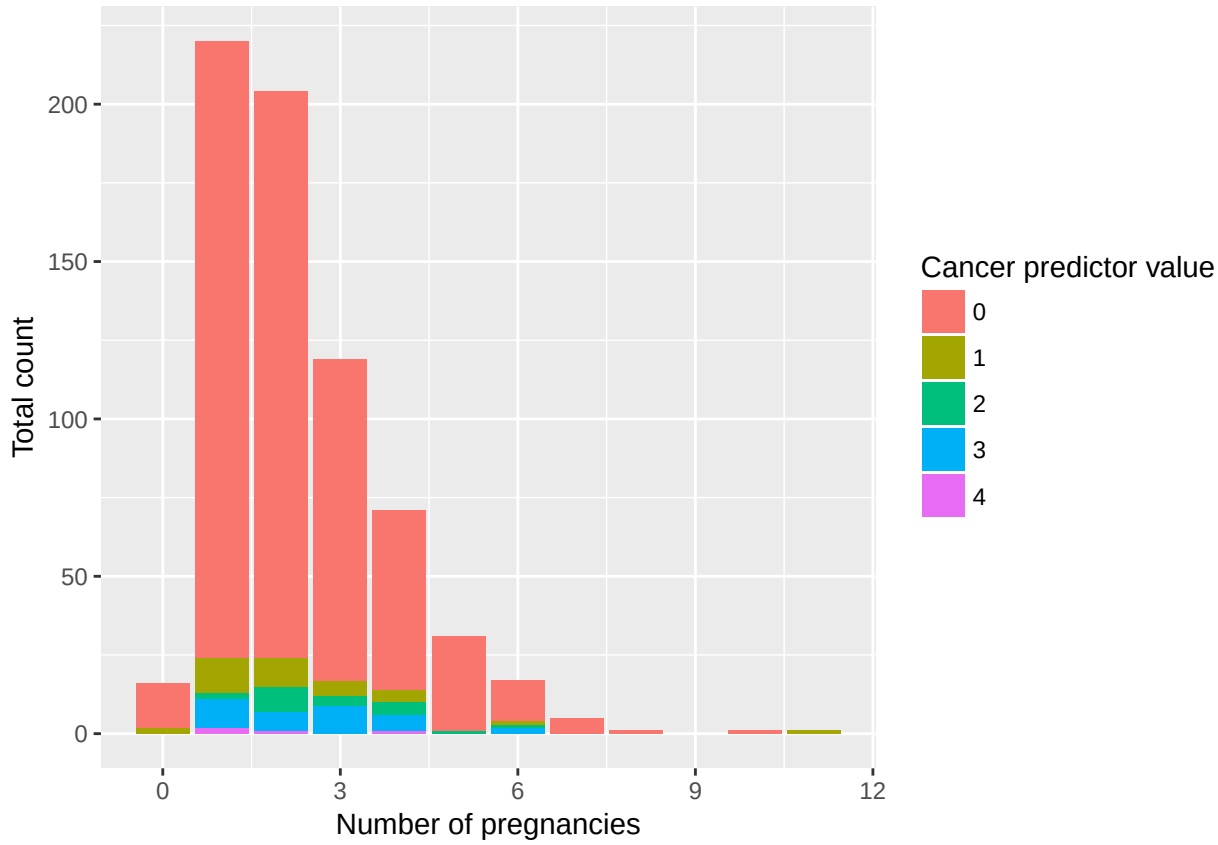
Since the human papilloma virus (HPV) is the main risk factor for cervical cancer it makes sense to investigate variables that speak to this. From the data provided, sexual activity indicators are most salient, as HPV is contracted via sexual activity with an infected person.



Slightly skewed, but not terribly so.

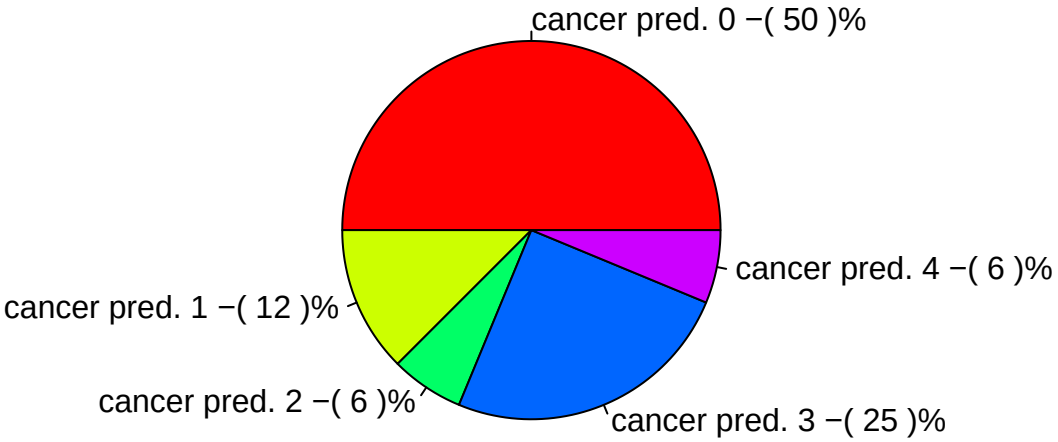


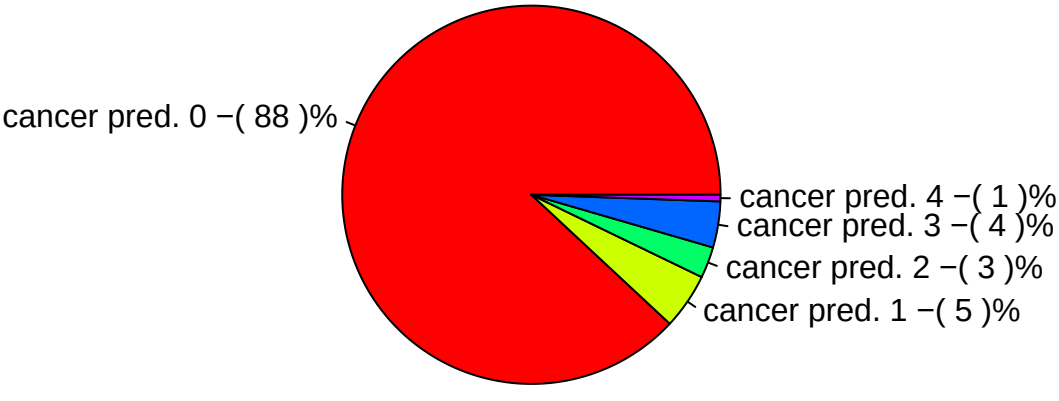
We can see that these values have a good normal distribution, and the dependent variable values seem to follow the data, indicating this may be an important factor in predicting the outcome.



Exhibits normal distribution, and dependent variable slightly follows the data.

HPV


Pie Chart of cancer predictor values for patients with HPV
cancer pred. 0 −(50)%
cancer pred. 4 −(6)%
cancer pred. 1 −(12)%
cancer pred. 2 −(6)%
cancer pred. 3 −(25)%

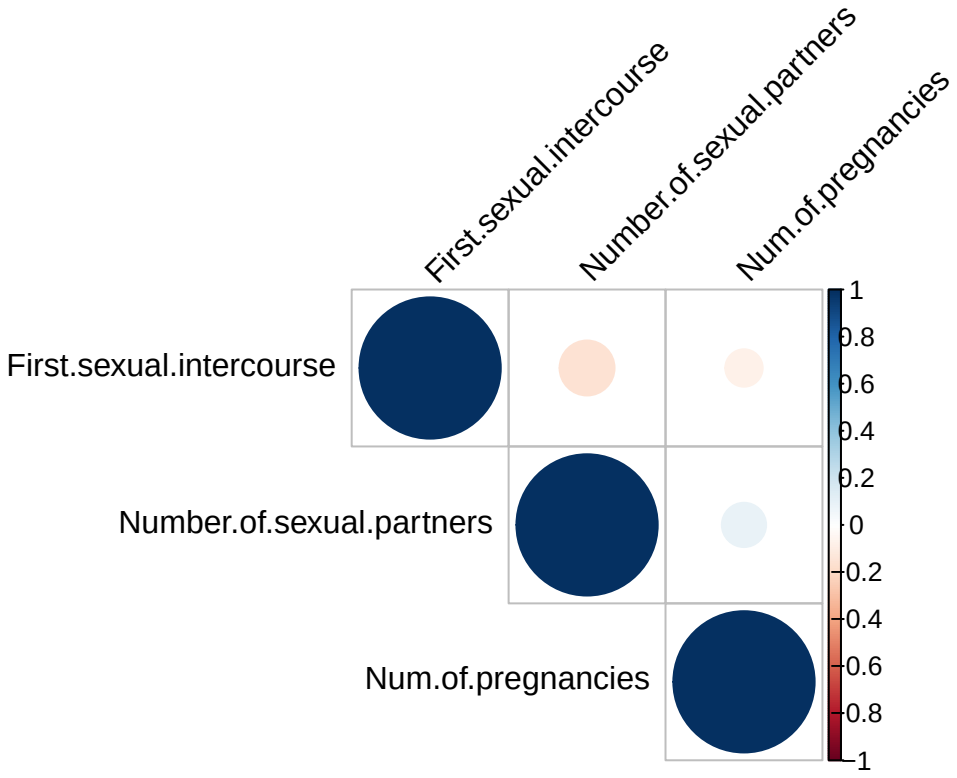

Pie Chart of cancer predictor values for patients without HPV
cancer pred. 0 −(88)%
cancer pred. 4 −(1)%
cancer pred. 3 −(4)%
cancer pred. 2 −(3)%
cancer pred. 1 −(5)%

These pie graphs confirm our review of literature, showing that patients with HPV show a higher prevalence of cervical cancer markers.

Multicollinearity across variables

Sexual history

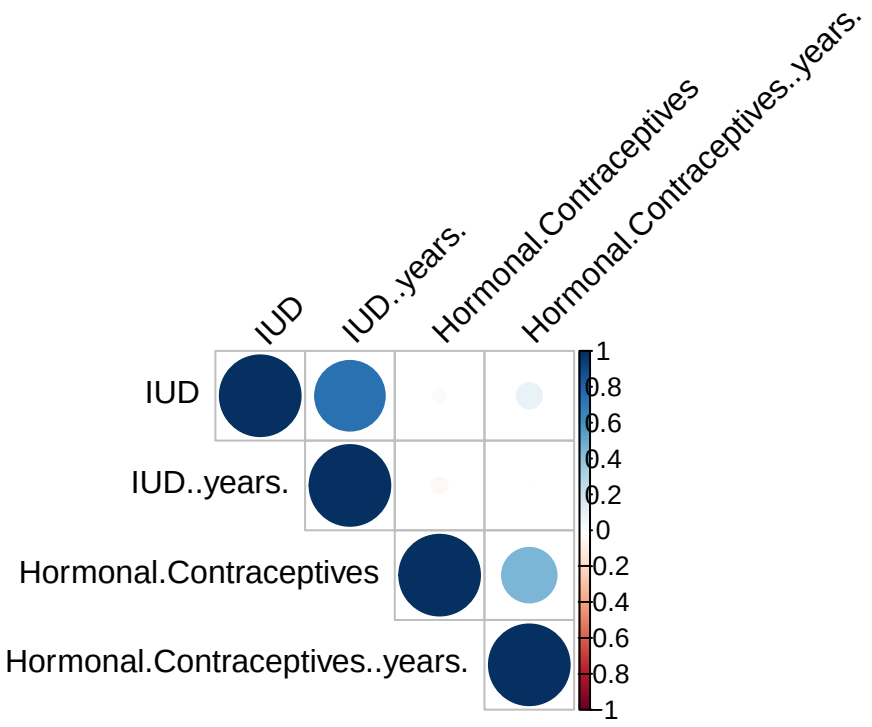
Captured in: `Number.of.sexual.partners`, `First.sexual.intercourse`, `Num.of.pregnancies`



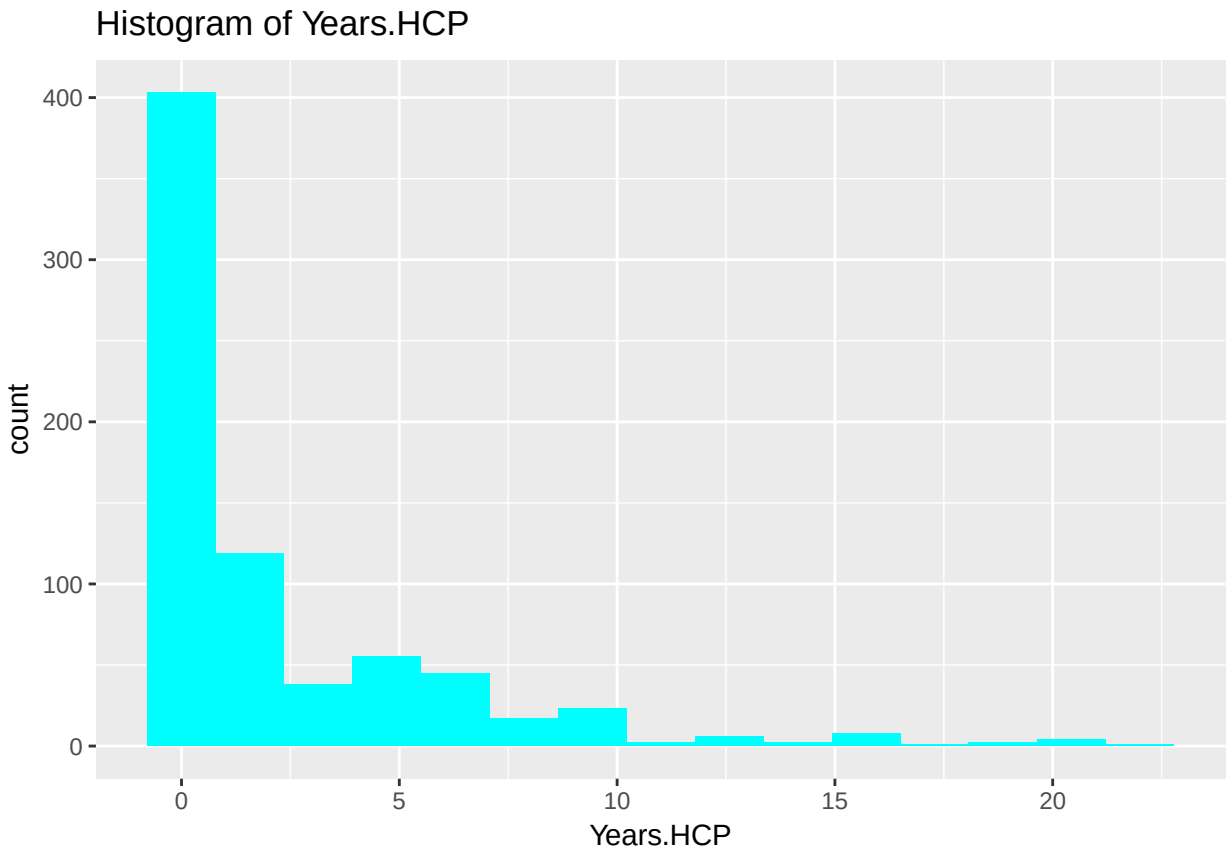
These variables are relatively uncorrelated, and thus can remain as they are.

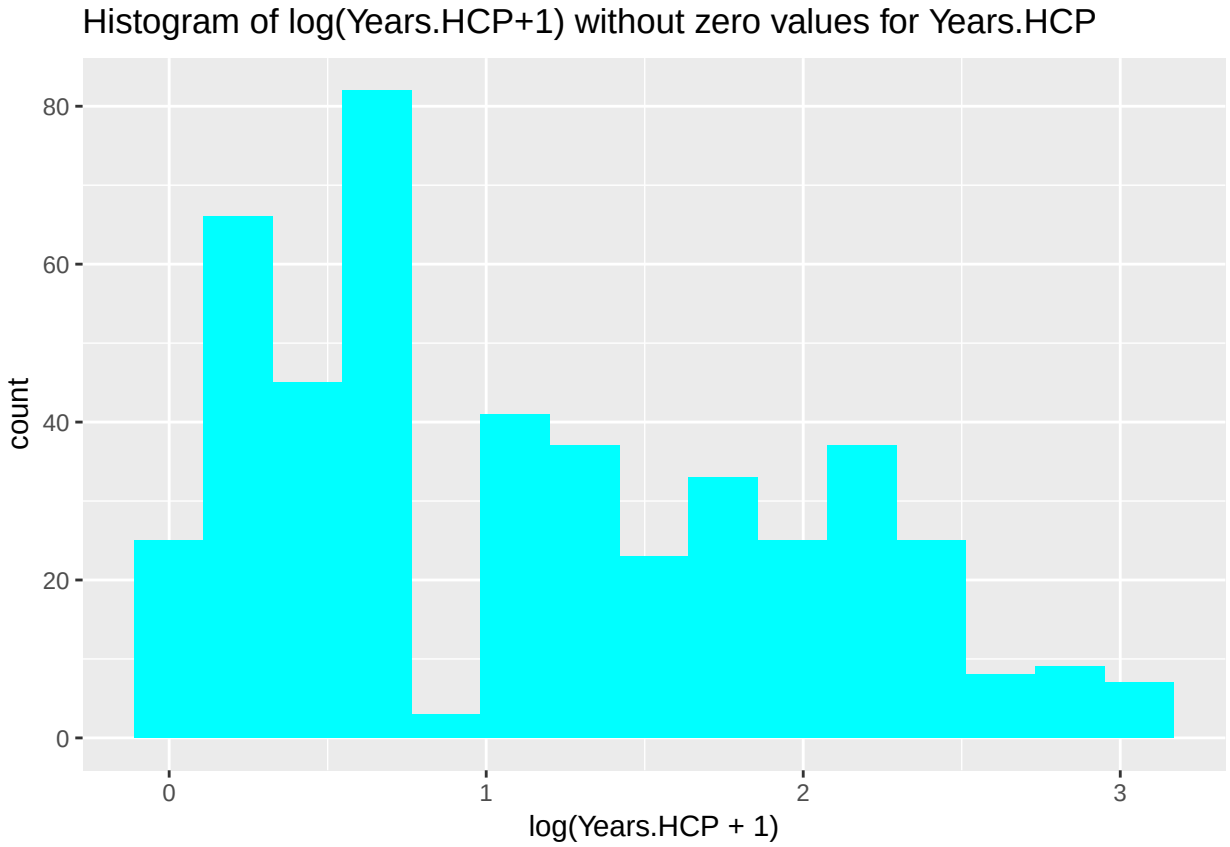
Contraception - Hormonal contraceptives & IUDs

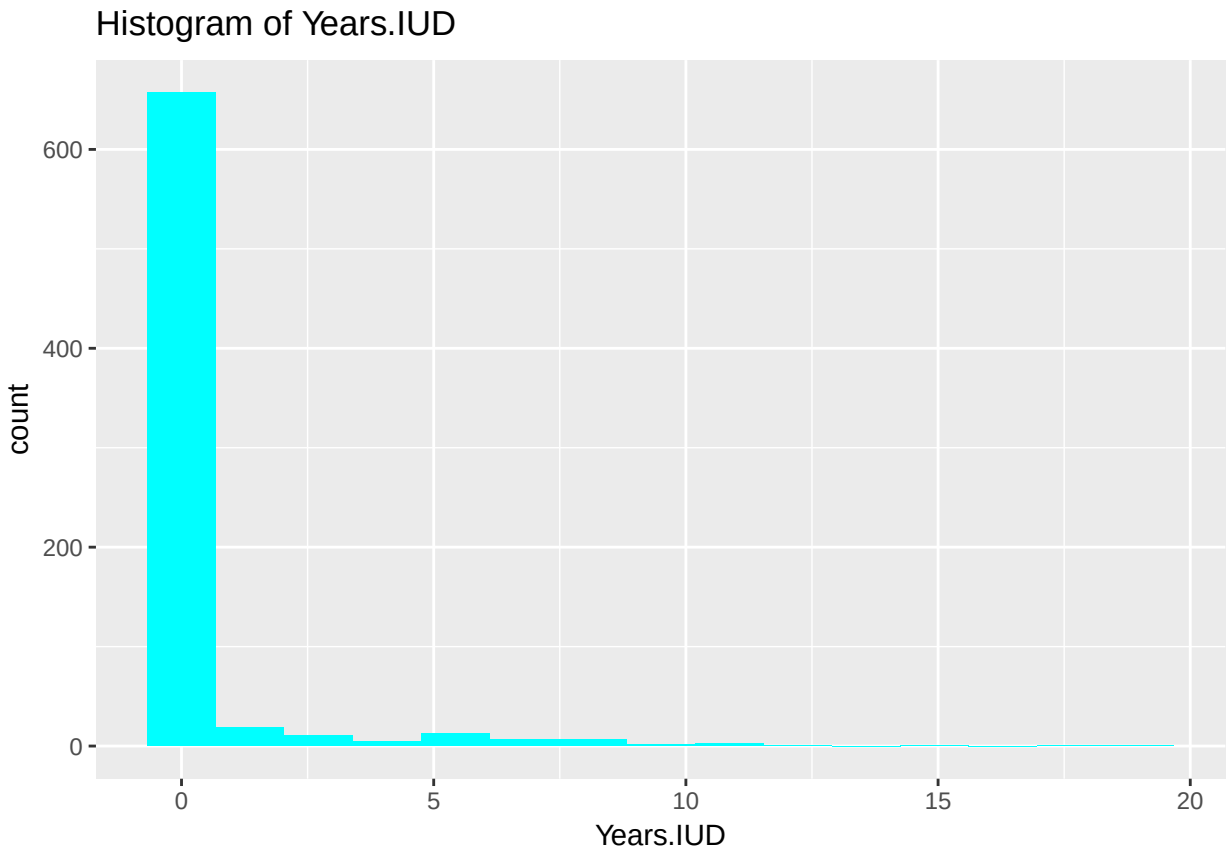
Captured in: `Hormonal.Contraceptives`, `Hormonal.Contraceptives..years.`, `IUD`, `IUD..years.`

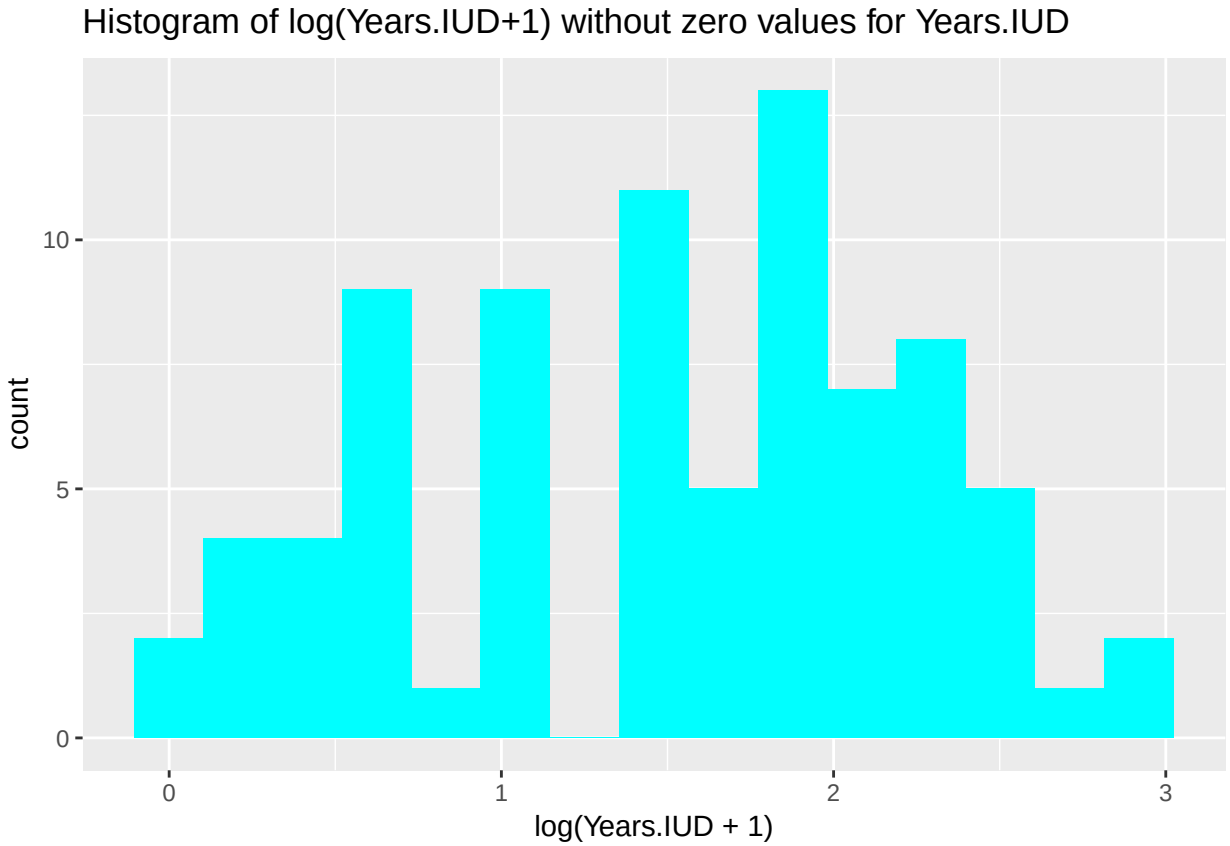


Results indicate that we should combine `IUD`/`IUD..years.`, and `Hormonal.Contraceptives`/`Hormonal.Contraceptives..yea`


Histogram of Years.HCP
count
400
300
200
100
0
0
5
10
15
20
Years.HCP


Histogram of log(Years.HCP+1) without zero values for Years.HCP
count
80
60
40
20
0
0
1
2
3
log(Years.HCP + 1)


Histogram of Years.IUD
count
600
400
200
0
0
5
10
15
20
Years.IUD

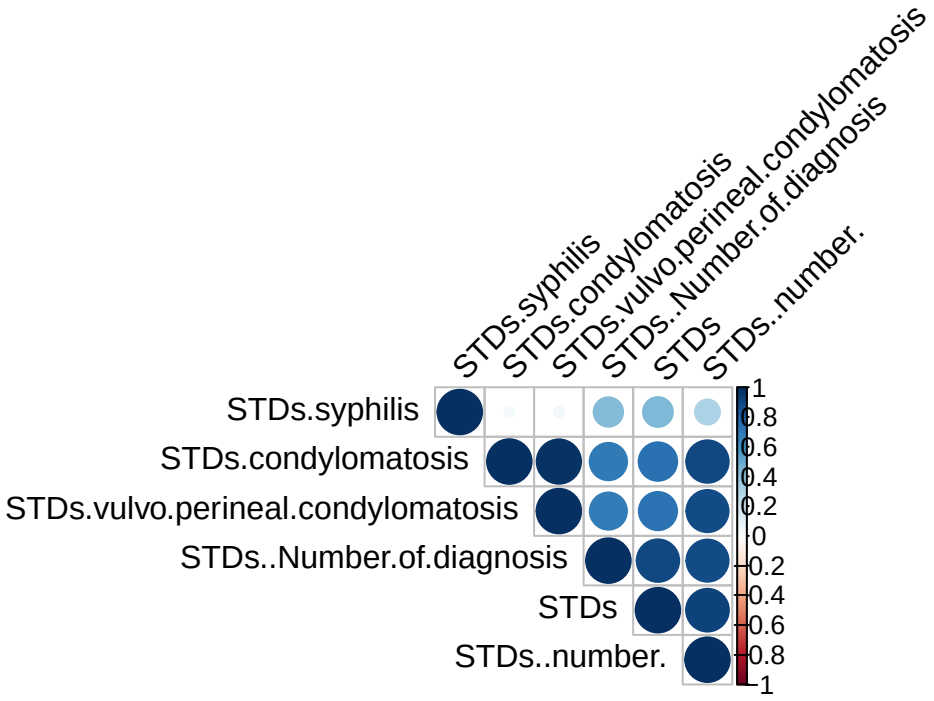


Similarly to `Smoker.status`, two new variables were created from the four original correlated variables. `log.Years.HCP` was created by multiplying $Years.HCP = Hormonal.Contraceptives(yes = 1, no = 0) \times Hormonal.Contraceptive.Years$ then taking the log transofrmation: $log.Years.HCP = log(Years.HCP + 1)$. The histograms before and after this process can be seen above. `log.Years.IUD` was created by multiplying $Years.IUD = IUD(yes = 1, no = 0) \times IUD.Years$, then taking the log transofrmation: $log.Years.IUD = log(Years.IUD + 1)$. As expected the histograms of the log transformed variables show a much better distribution than the originals.

The preliminary variables `Hormonal.Contraceptives`, `Hormonal.Contraceptives..years.`, `IUD`, `IUD..years.`, `Years.HCP`, and `Years.IUD` were then removed from the dataframe in lieu of the log values.

STIs

Captured in: `STDs`, `STDs..number.`, `STDs.condylomatosis`, `STDs.vulvo.perineal.condylomatosis`, `STDs.syphilis`, `STDs..Number.of.diagnosis`



Unsurprsingly, there exists high correlation between many of these variables. To address this, we remove those with high correlation. As the figure shows, all variables except `STDs.syphilis` are highly correlated with `STDs.condylomatosis`. Consequently we leave those that are uncorrelated (`STDs.syphilis`, and `STDs.condylomatosis`) in the model, and remove the rest.

Final dataset, Testing & Training Data Split

In this study, we classify a patient as having low risk of developing cerivical cancer if cancerPred is 0. If cancerPred is 1-4, the patient is considered to have high risk of developing cerivical cancer.

```
data1$cancerPred <- ifelse(data1$cancerPred>0, 1, 0)
data1$cancerPred <- as.factor(data1$cancerPred)
```

Finally, we split our data into training and testing with 30% of data set aside as testing set.

```
# Split into training and testing
set.seed(1)
index <- sample(nrow(data1), round(nrow(data1) * 0.7))
train <- data1[index,]
test <- data1[-index,]
```

We're now ready to model.

Modelling

We developed two models to predict the risk of cervical cancer: 1) Logistic Regression with LASSO and 2) Random Forest. We introduce each and proceed to compare them with confusion matrix metrics and AUC.

Logistic Regression with LASSO

We conduct LASSO with 10-fold cross validation on the training set to compare the performance of the logistic regression.

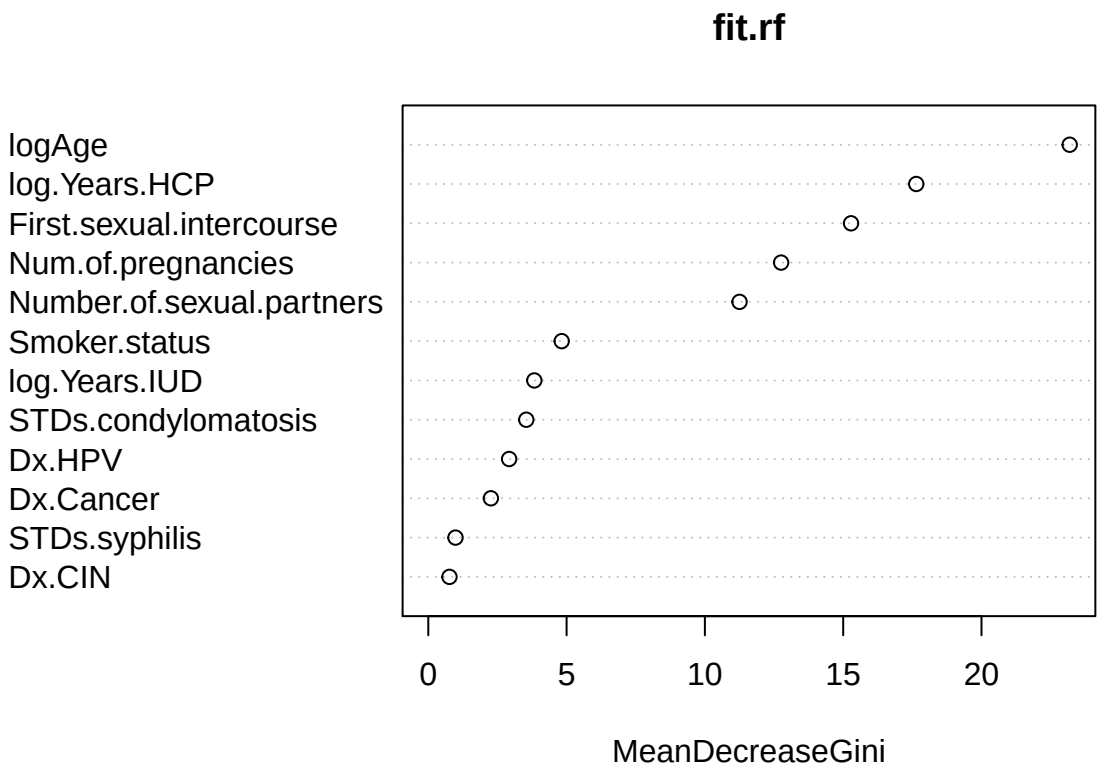
```
Analysis of Deviance Table (Type II tests)

Response: cancerPred
                    LR Chisq Df Pr(>Chisq)
STDs.condylomatosis    5.385  1  0.0203102 *
Dx.HPV                12.513  1  0.0004041 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This model shows that `Dx.Cancer` and `STDs.condylomatosis` are significant factors at the 0.05 level in predicting cervical cancer risk.

Random Forest

The random forest model deems `logAge`, `log.Years.HCP`, `First.sexual.intercourse`, `Num.of.pregnancies`, and `Number.of.sexual.parters` as important factors in predicting cervical cancer risk. These variables all conform to our original hypotheses.

```
                          MeanDecreaseGini
Number.of.sexual.partners       11.2534052
First.sexual.intercourse        15.2843367
Num.of.pregnancies              12.7560783
STDs.condylomatosis              3.5429690
STDs.syphilis                    0.9836865
Dx.Cancer                        2.2629130
Dx.CIN                           0.7663258
Dx.HPV                           2.9238029
Smoker.status                    4.8236531
logAge                          23.1889062
log.Years.HCP                   17.6418464
log.Years.IUD                    3.8331406
```



Comparison of Methods

We will compare the models with confusion matrix metrics and area under curve.

Confusion Matrix

We apply our two models to the test data and see how well they predict cancer risk. In classifying our probabilities, one might set phat to be 0.5. However, since this is a medical prediction, we want to data to be more sensitive in order to decrease the chance of false negatives - we wouldn't want to tell a patient who has cancer he/she has does not have cancer. Given that, we set phat to be 0.25 for class assignments.

The confusion matrix reports the following information:

```
fit1.2pred   0   1
         0 172  15
         1   9   4
```

The logistic regression model returns MCE of 0.12 on our test set. It has sensitivity of 0.2105263 and specificity of 0.9502762.

Let's compare these numbers to that of the random forest model:

```
fit2.pred   0   1
        0 153  12
        1  28   7
```

The random forest model returns MCE of 0.2 on our test set. It has sensitivity of 0.3684211 and specificity of 0.8453039.

We summarize the performance of the two models as follows:
- **Misclassification rate**: Logistic regression 0.12; Random Forest 0.20. Logistic regression has a lower misclassification rate.
- **Sensitivity**: Logistic regression 0.21; Random Forest 0.37. Random Forest has higher sensitivity, and thus has lower false negative rates.
- **Specificity**: Logistic regression 0.95; Random Forest 0.85. Logistic regression has higher specificity, and thus have lower false positive rates.

If our objective is to minimize misclassification rate, logistic regression model seems to perform better. However, if we want to minimize false negative rates (ie. minimize errors in diagnosing a patient with cancer as without), we'd be better off with the random forest model.

ROCs and AUC

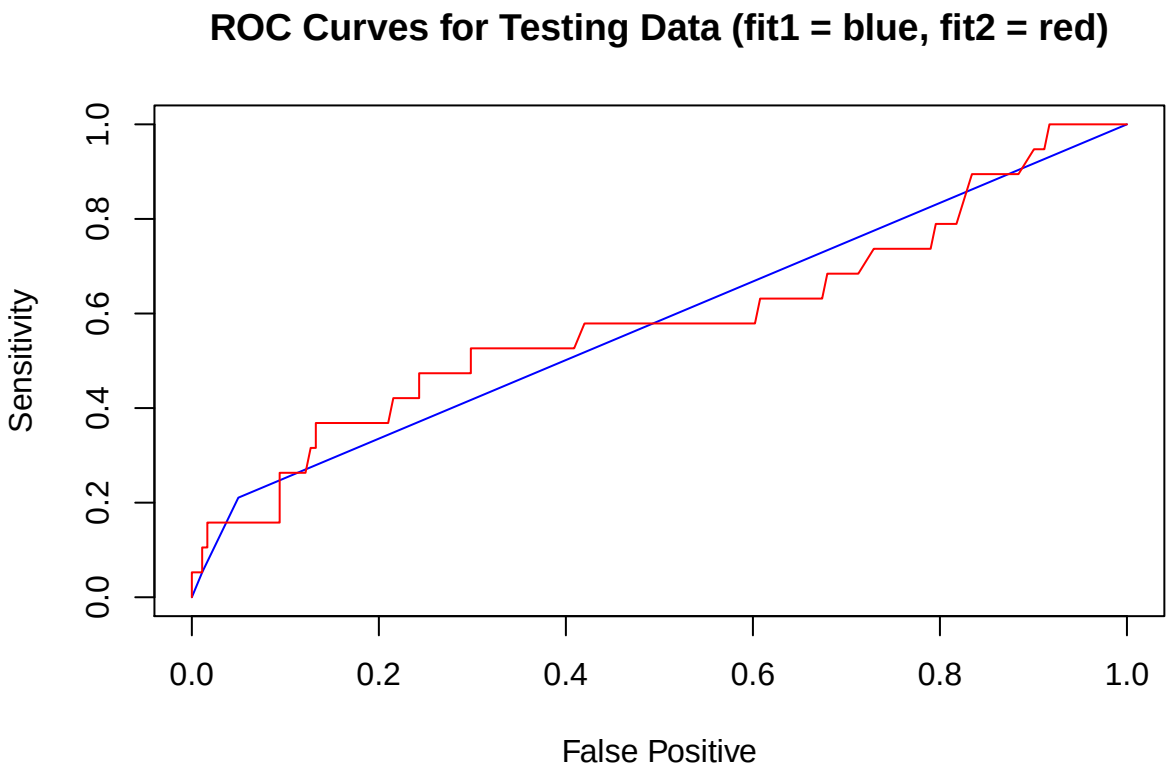
Let's look at ROC and area under curve to further compare the models.


```
Area under the curve: 0.5824
```

```
Area under the curve: 0.9996
```

The plot below shows that random forest has higher auc than logistic regression model at auc of 0.9995906. In fact, at 0.5823501, logistic regression model's auc is so low that it is not much better than guessing.

Interestingly, the picture is very different when we plot ROC with our test set.




```
Area under the curve: 0.5805
```

```
Area under the curve: 0.5845
```

Neither model performs particularly well on the test set (logistic regression auc 0.5805467, random forest auc 0.5844722. This may be due to overfitting of the training set. We will discuss this further in the next section (Validity of Results).

Validity of Results and Future Improvement

Our model predicts `Dx.HPV` and `STDs.condylomatosis` as being the most important variables. Based on literature review, it makes sense that a diagnosis HPV is the strongest predictor of whether or not somebody is at high risk for contracting cervical cancer.`STDs.condylomatosis` does not make as much sense, since condyloma or genital warts are generally caused by HPV strains that do NOT lead to cancer. However, in the initial EDA, we chose to include `STDs.condylomatosis` over some of the other highly correlated variables that captured sexual activity. Thus, a possible explanation for this effect is that `STDs.condylomatosis` is a confounding variable and that the variable truly responsible for the effect is `STDs..number.`.

Unfortunately, our model did not reveal the importance of known risk factors for cervical cancer (such as age at initial intercourse, number of sexual partners, patient age, CIN, among others). This could be due to two possibilities: First, it could be that the manipulation in the EDA ended up removing some of the variability that would allow us to capture these effects. Second and most likely, it coud be that these known risk factors have relatively small effects when compared to the #1 risk factor and only contibute minimally to the overall risk. Thus, our study might have been underpowered for the purposes of identifying cervical cancer predictors.

Future improvements to our model should consider an increased sample size (particularly trying to enroll more patients whose `cancerPred` score is "high risk" 1-4), which would increase the power of the study and our ability to detect small effects. Another improvement could be the use of a more definitive cancer diagnosis variable, such as the presence of an active cervical cancer diagnosis code. Lastly, it would be interesting to explore the effects of HPV vaccination by including variables such as whether or not somebody received Gardasil or Cevarix, how many doses were administered, and at what ages at which the doses were given.

Conclusion

In identifiying factors that influence cervical cancer risk, we employed two models (logistic regression and random forest). We compared the two models with sensitivity, specificity, false nagative rate, false positive rate, unweighted misclassification errors (MCE), ROC curves and AUC. We find that random forest outperforms logistic regression in classifying the training data (with an AUC of near 1); however, both models had similar performance in terms of AUC for the test data (both around ~0.58). Also, while logistic regression outperformed random forest in terms of MCE for the specified rule, random forest had a higher sensitivity ($P(\hat{Y} = 1|Y = 1)$) and lower false negative rate ($P(\hat{Y} = 0|Y = 1)$), which are two desirable qualities when classifying potential cancer patients. Therefore, between the two models, we recommend employing the random forest model for future predictions. Note however that while the random forest model has the upper hand in this analysis, there is still significant room for improvement in data collection to improve the model's prediction powers. Future directions for improvement include significantly increasing our sample size, particularly increasing the proportion of "high risk" patients, and exploring the effect of vaccination against HPV.

Appendix

Original variable descriptions

Demographic

- `Age` - (int) Age
- `Smokes` - (bool) Smokes
- `Smokes..years.` - Smokes (years)
- `Smokes..packs.year.` - Smokes (packs/year)

Contraception

- `Hormonal.Contraceptives` - (bool) Hormonal Contraceptives
- `Hormonal.Contraceptives..years.` - (int) Hormonal Contraceptives (years)
- `IUD` - (bool) IUD (Intrauterine device)
- `IUD..years.` - (int) IUD (years)

Sexual activity history

- `Number.of.sexual.partners` - (int) Number of sexual partners
- `First.sexual.intercourse` - (int) First sexual intercourse (age)
- `Num.of.pregnancies` - (int) Num of pregnancies

Sexual health history

- `STDs` - (bool) STDs
- `STDs..number.` - (int) STDs (number)
- `STDs.condylomatosis` - (bool) STDs:condylomatosis
- `STDs.cervical.condylomatosis` - (bool) STDs:cervical condylomatosis
- `STDs.vaginal.condylomatosis` - (bool) STDs:vaginal condylomatosis
- `STDs.vulvo.perineal.condylomatosis` - (bool) STDs:vulvo-perineal condylomatosis
- `STDs.syphilis` - (bool) STDs:syphilis
- `STDs.pelvic.inflammatory.disease` - (bool) STDs:pelvic inflammatory disease
- `STDs.genital.herpes` - (bool) STDs:genital herpes
- `STDs.molluscum.contagiosum` - (bool) STDs:molluscum contagiosum
- `STDs.AIDS` - (bool) STDs:AIDS
- `STDs.HIV` - (bool) STDs:HIV
- `STDs.Hepatitis.B` - (bool) STDs:Hepatitis B
- `STDs.HPV` - (bool) STDs:HPV
- `STDs..Number.of.diagnosis` - (int) STDs: Number of diagnosis
- `STDs..Time.since.first.diagnosis` - (int) STDs: Time since first diagnosis
- `STDs..Time.since.last.diagnosis` - (int) STDs: Time since last diagnosis

Medical history

- `Dx.Cancer` - (bool) Dx:Cancer (person had previous cervical cancer diagnostic)
- `Dx.CIN` - (bool) Dx:CIN (person had previous diagnostic of Cervical intraepithelial neoplasia - the potentially premalignant transformation and abnormal growth (dysplasia) of squamous cells on the surface of the cervix)
- `Dx.HPV` - (bool) Dx:HPV
- `Dx` - (bool) Dx
- `Hinselmann` - (bool) Hinselmann: target variable
- `Schiller` - (bool) Schiller: target variable
- `Citology` - (bool) Cytology: target variable
- `Biopsy` - (bool) Biopsy: target variable