Data Analysis of Cervical Cancer

Andrew Wang

9/28/2020

Introduction

In this report, we will be analyzing a set of cervical cancer data, and discuss what it teaches us about cervical cancer. First, we will go over background information about cervical cancer, HPV, and the data set at hand. After this background information, we'll use R to conduct some data analysis regarding trends and patterns in our dataset.

To start off with, cervical cancer is a cancer which comes from the cervix, a cylinder-shaped tissue connecting the uterus to the vagina. This type of cancer is only present in women. Like other cancers, it is due to the uncontrolled growth of mutated cells, which invade other parts of the body to spread. Human Papillomavirus Infection (HPV) is responsible for more than 90% of cases. However, most people who have HPV do not developed cervical cancer. Furthermore, development of cervical cancer also revolves around a host of other lifestyle factors, including smoking, having many sexual partners, having a weak immune system, and many others.

To identify cervical cancer, screening is conducted using Pap tests or acetic acid. Treatment of cervical cancer may include surgery, chemotherapy, radiation therapy, and other treatments. The current five-year survival rate, meaning the percentage of people alive five years after they are dianosed with the disease, is 68% in the United States. However, a lot of the treatment procedure revolves around how early you can find it. Furthermore, HPV plays a large role in cervical cancer, and recognition/testing for HPV will also be an important forewarning for potential cervical cancer. Thus, some background knowledge of HPV will be explained as well.

Human Papillomavirus Infection

Human Papillomavirus Infection (HPV) is a non-enveloped DNA virus, specifically evolved to infect human epithelial cells. Epithelial cells are the cells which line outer surfaces of organs and blood vessels, serving as a barrier between parts of the body. There are over 100 types of HPV, and they are categorized by the specific epithelial cells which they target. For example, some may target epithelial cells on the skin, mucous membranes of the respiratory tract, or anal and genital regions. They can cause benign tumors and in some cases lead to cancer. The current HPV vaccines protect against two-to-seven of the deadliest strains and prevent many potential cases of cervical cancer.

Causes

HPV is spread through contact with infected epithelial cells. Activities which can increase the risk of HPV exposure include things like having multiple sexual partners or delivering a baby through an infected birth canal. Furthermore, HPV is more likely to infect those with already weak immune systems. Luckily, its transformation into cancer is not guarunteed, it depends on the strain of HPV, along with lifestyle factors like tobacco usage and radiation exposure.

Symptoms

Symptoms of HPV vary depending on the strain of the virus, but most strains share some similar characteristics. Among those include warts of various kinds, such as benign warts, plantar warts, and flat warts. Infection in the respiratory system can cause things like voice changes or high-pitched breathing, particularly if the infection in on the larynx. Specifically, HPV types 6 and 11 cause most larynx infections and genital warts, but these are considered low risk strains, and usually have no progression beyond warts. However, this is not the case for HPV strains 16 and 18, which have a high risk of transforming into various kinds of cancers.

Treatment

Treatment of HPV can be any combination of various methods. To remove warts, you could use salicylic acid, liquid nitrogen, or laser/surgical removal. Immune modifiers may be used to boost immune capabilities in fighting the virus in the future, but oftentimes no treatment is necessary as many infections will resolve themselves over time, particularly if you are young. As with most other communicable diseases, prevention is the best course of action. Simple things like limiting contact with infected people or recieving the vaccine can save people a lot of potential trouble.

Epithelial Cells - how are they infected by HPV?

As stated in the previous section, epithelial cells line the outer surfaces of organs and blood vessels, separating the interior of the body. They serve as a protective barrier against infectious microbes. In many cases, there are multiple layers of epithelial cells, starting the basal cells at the base. The basal cells are responsible for replenishing all epithelial cells in their region, and their replication pushes older epithelial cells further out, forming a protective layer. These older epithelial cells become more oval-shaped, and are eventually shed when they become too old.

While basal epithelial cells are usually well protected by the more mature epithelial cells on the outside, lesions can allow the HPV virus to get into the body and infect the basal cells. Once in the body, HPV can replicate with or without insertion into the basal cell's DNA. The HPV virus can alter proteins of the viral genes E6 and E7. The proteins of these two viral genes control the replication patterns of the epithelial cells. By altering the proteins which prevent unregulated growth, HPV can cause uncontrolled cell replication. This uncontrolled growth forms warts.

In the following sections, I will use R to conduct data analysis on the various factors of cervical cancer and HPV.

Analysis with R

Data cleanup and setup

In this section, we are going over various functions and methods for cleaning data in R, which will allow us to better perform analysis later on it. First, we need to clean up and set up the R environment with the working directory, and we import an R script for reference later. Kee in mind that your directory structure is likely different from mine, and adjust accordingly.

```
# clean up and setup
rm(list=ls()) # clean up any old stuff in R
setwd("C:/Users/hyper/OneDrive/Desktop/Desktop Folders/Programming/R/Assignments/Week
 5") # go to this folder
#load up myfunctions.R
source("C:/Users/hyper/OneDrive/Desktop/Desktop Folders/Programming/R/myfunctions.R")
```

Load the library

Next, we want to load the library we'll be using to clean and prepare the data. The library name is `tidyverse`. Make sure you install the library first, which I will not be doing here as I already have it installed.

```
#library import
library(tidyverse)
```

Importing the csv file

After installing the library, it is time to import the csv file. First, make sure your csv file is in the same location as your specified working directory (defined above). This will make sure R knows where to go to find the csv file. After this, we run a few initial viewing functions to look at basic attributes of our data. In the code below, I've commented out the `str` and `summary` function calls, as they made the paper five pages longer. Make sure to give the dataset an appropriate name. Our dataset is called `cancerData`, as it contains data about cervical cancer. In the last line of code, I simply sort the rows in descending order according to age. While this is not needed, I thought it would prove more pleasing to view.

```
#input csv file
cancerData <- read.csv("cervicalCAClean.csv")
dim(cancerData) #gives dimensions of data
```

```
## [1] 858  49
```

```
#str(cancerData) #structure overview
#summary(cancerData) #stat info on numeric variables
View(cancerData) #view the data as a table
cancerData <- arrange(cancerData,age)
```

While we would normally spend some time to clean empty values out of the data. that has already been done. If you want to verify, you can use `print(sum(is.na(cancerData)))`, which should return 0.

Data munging

After initial impression about the dataset, I've chosen a few columns to remove. I am removing these columns because I feel that their meaning is lost, and would actually hinder any analysis done on the data rather than help it. However, it is important to be very careful in removing columns, once you remove a column you can't get it back. Only remove columns in very select circumstances.

```
# delete useless columns
cancerData$X <- NULL
cancerData$dx <- NULL
```

Determining influential factors for analysis

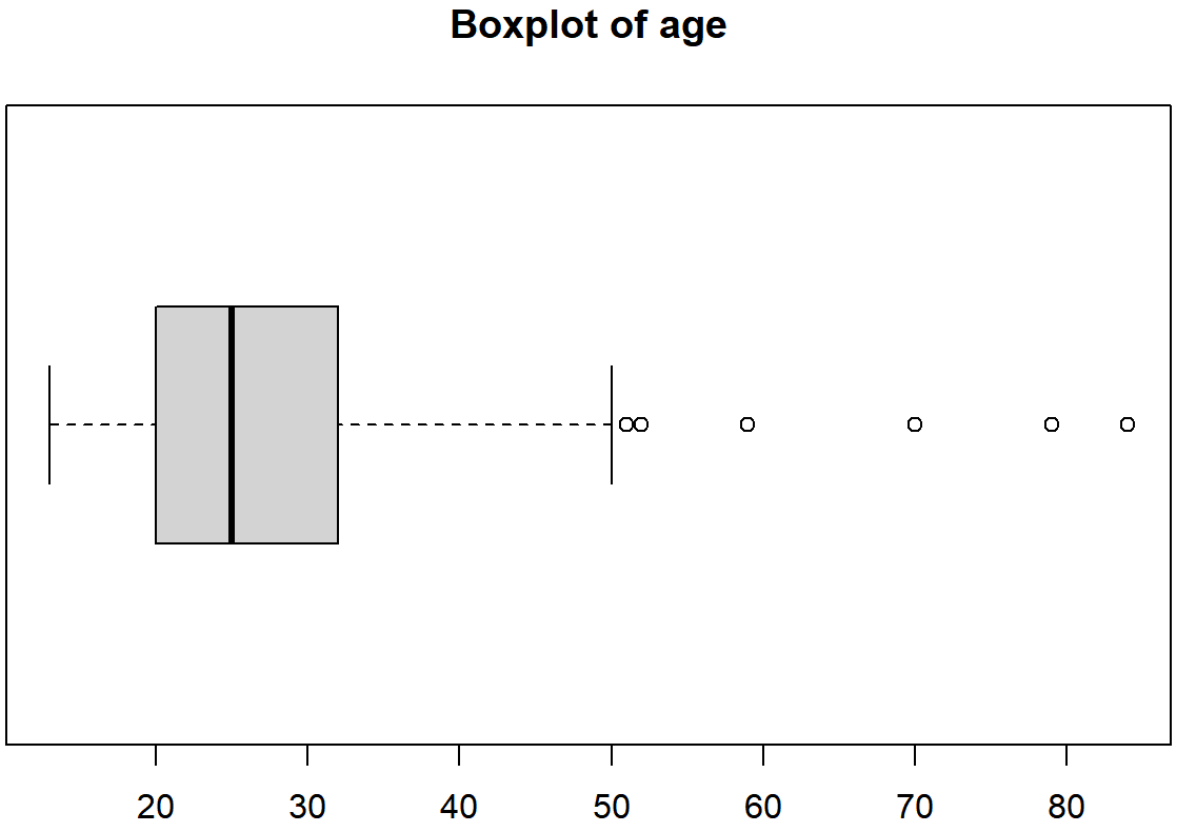
Our dataset is very large, so it is critical that we narrow it down to a handful of factors that we want to analyze. For this paper, I have chosen the age, number of sexual partners, total cigarette packs smoked, number of pregnancies, and number of STDs to evaluate. Note that there is not total cigarette column (yet) in the dataset. I thought that neither the years smoked or packs per year were good enough for solid analysis, as there could be misleading conclusions drawn from either of them. For example, someone who has smoked for many years, but only smokes one pack of cigarettes per year, or someone who has smoked for only one year, but has smoked many packs. Therefore, I thought it was necessary to add a new column, which I called `total.smokes`, to show the total number of cigarette packs smoked.

```
#add new column for analysis
cancerData <- mutate(cancerData, total.smokes = cancerData$smokes.years * cancerData$sm
okes.packs.year)
```

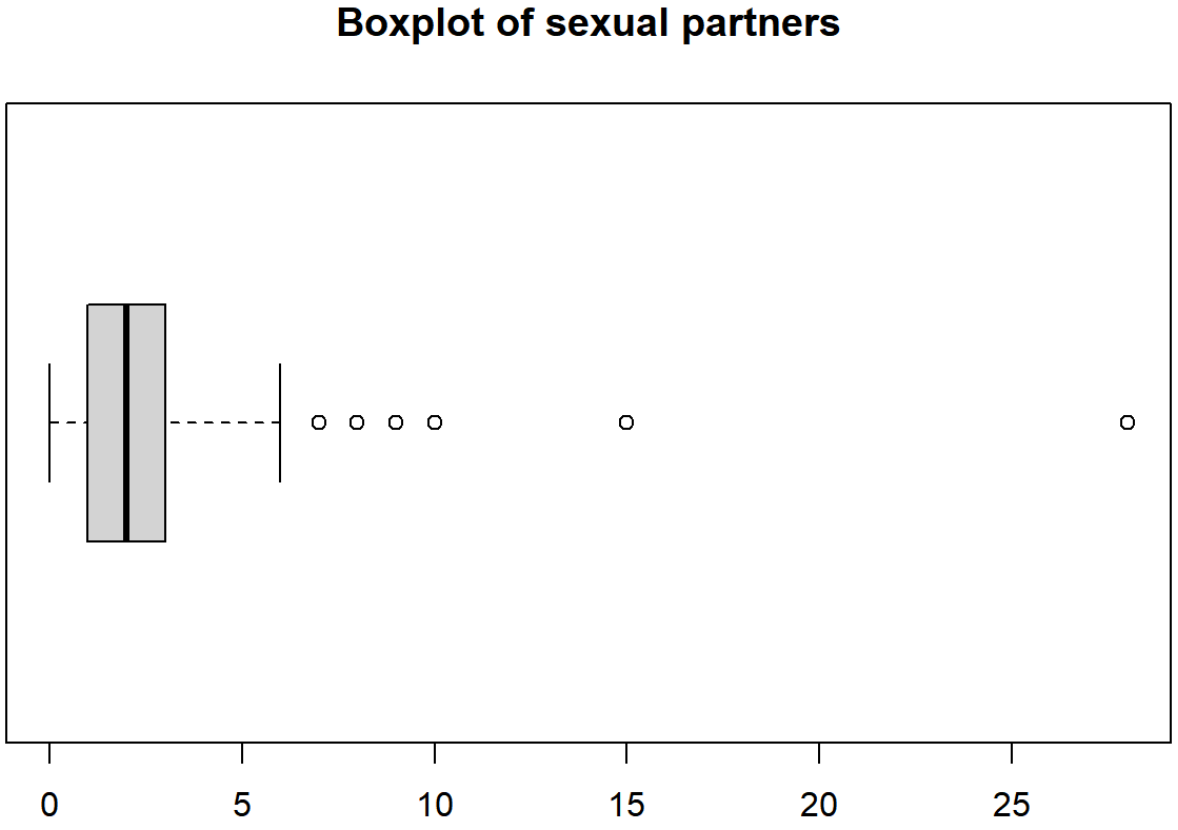
Basic visual analysis of influential factors

Next, we want to conduct a basic visual analysis of the influential factors, so we can get a grasp for what their distribution looks like. I've decided to use boxplots as the visualization tool, as histograms were lopsided and thrown off by extreme outliers. To go over the code, first we set up a graphics window so that the boxplots don't take as much space. Next, we repeat the same boxplot code, basically just defining the column to make a boxplot of and giving it a name. Analysis follows after the code.

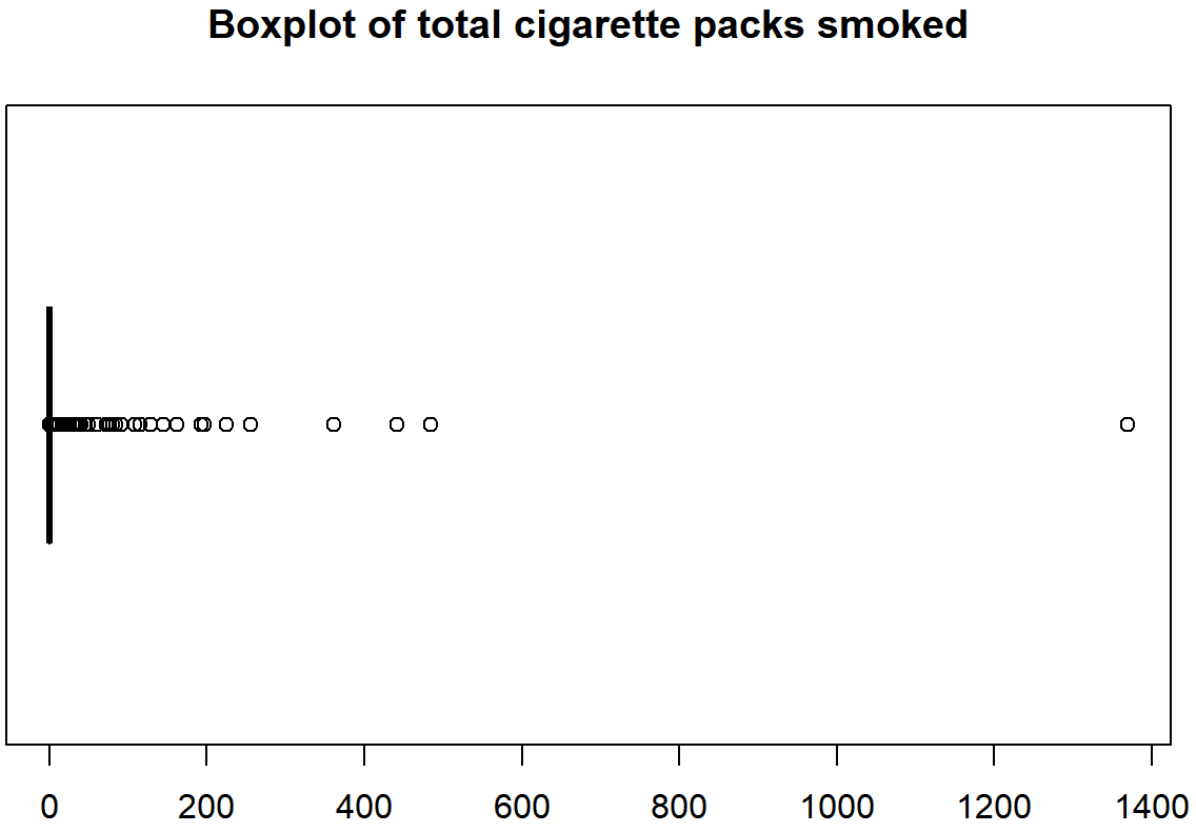
```
# Boxplot plots of various influencing factors (age, sex partners, total smokes, pregna
ncies, std amount)
#par(mfrow = c(3,2)) #setting up a graphics grid
boxplot(cancerData$age, main = "Boxplot of age", horizontal = TRUE)
```



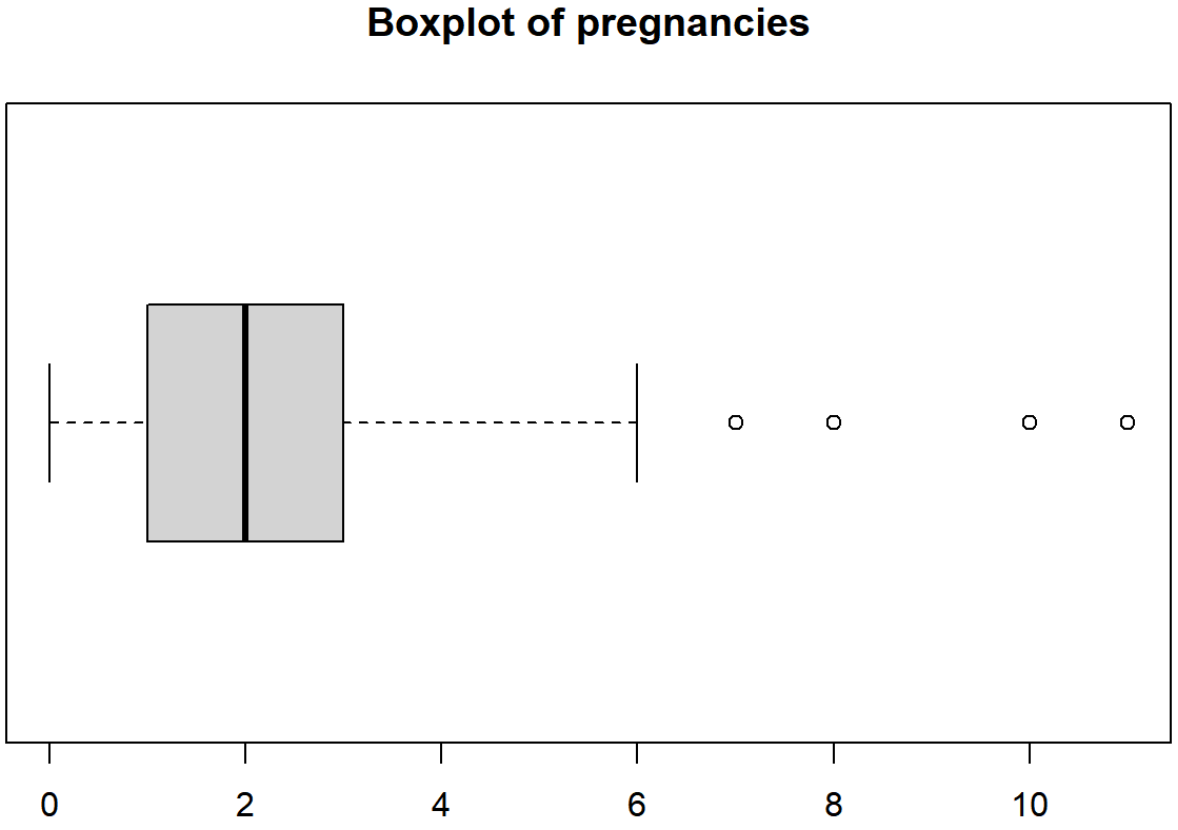

```
boxplot(cancerData$number.of.sexual.partners, main = "Boxplot of sexual partners", hori
zontal = TRUE)
```



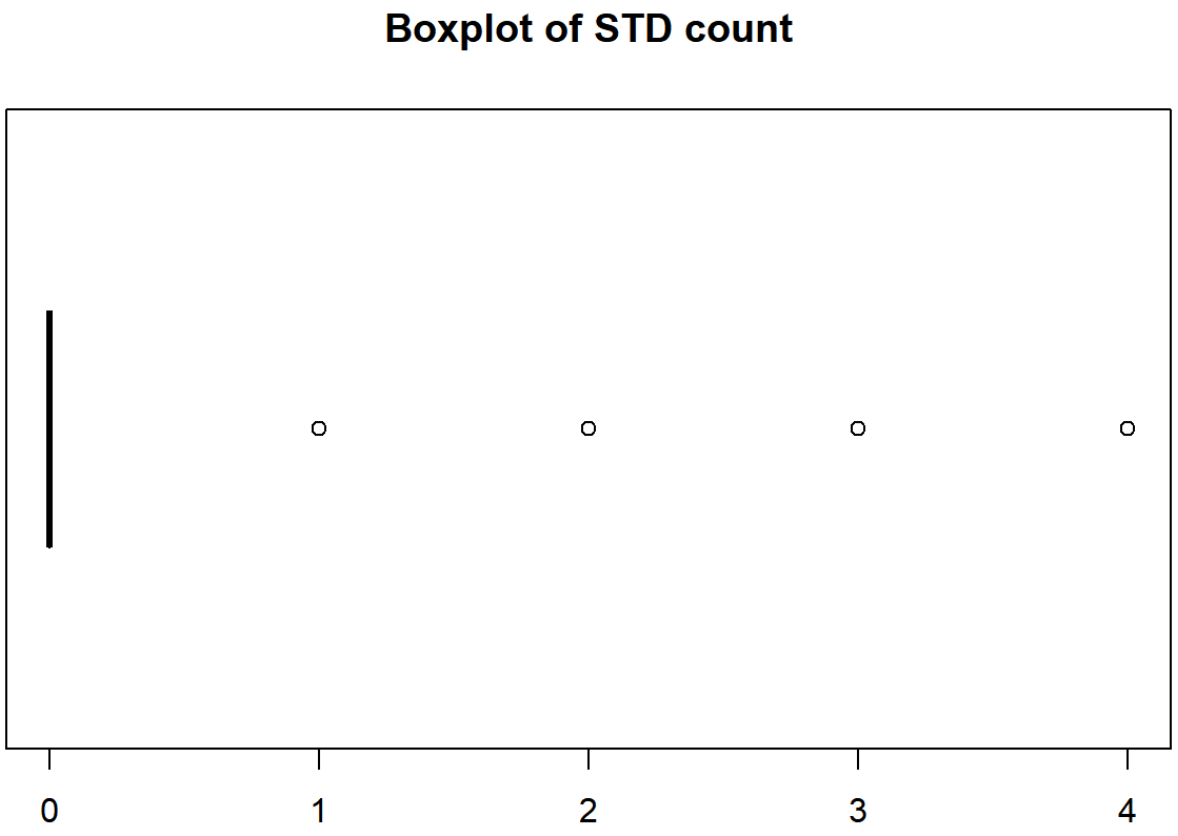

```
boxplot(cancerData$total.smokes, main = "Boxplot of total cigarette packs smoked", hori
zontal = TRUE)
```




```
boxplot(cancerData$num.of.pregnancies, main = "Boxplot of pregnancies", horizontal = TR
UE)
```




```
boxplot(cancerData$stds.number, main = "Boxplot of STD count", horizontal = TRUE)
```



Age The age boxplot is rather left-leaning, with a median around 23, and an IQR of around 20-35. It contains a few outliers to the right, but overall the data is of rather young people.

Sexual partners The sexual partner boxplot is also left-leaning, with a median around 2 and an IQR of similar range. This means most the data is closely clustered around the median. Like the age boxplot, there are a few outliers, with one extreme outlier.

Total cigarette packs The total cigarette packs smoked boxplot is extremely left-leaning, with a median around 0 and an IQR that is not even visible. This suggests that the data is extremely closely clustered around the median. However, there are many more outliers in this boxplot, most of them ranging from 0-200, but with one extreme outlier at 1369.

Pregnancies The total pregnancy boxplot, like most other box plots, is rather left leaning. It has a median of 2 with an IQR range of 1-3. There are few outliers.

STDs Finally, the STD count boxplot is the most simple of them all, with a median at 0 and an IQR that is not visible. There are a few outliers, but it means most people in the dataset have 0 STDs.

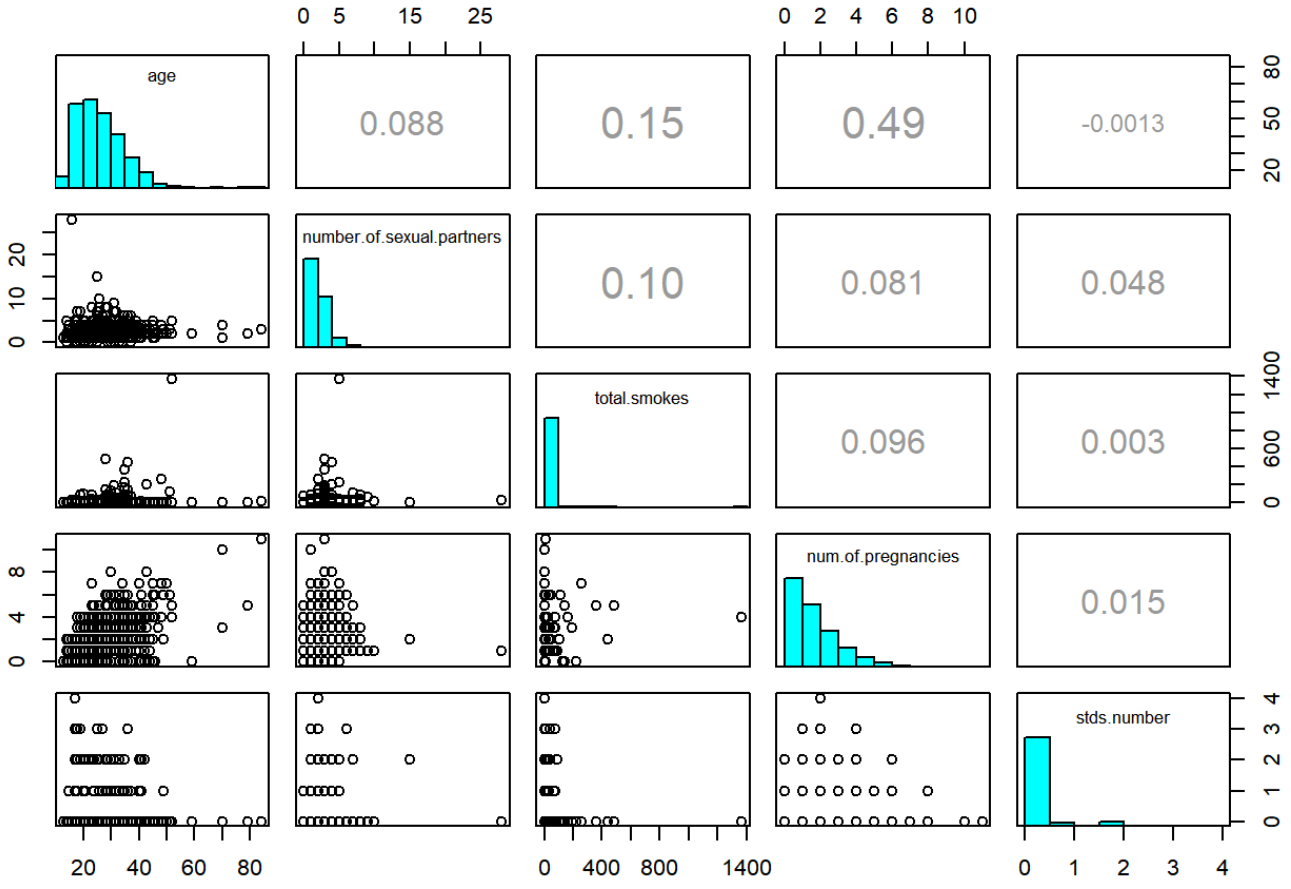
Pair correlation plot

Here, I want to evaluate the relationship between our influencing variables, so I'm using the `pair` function to do so. First, I set up a data subset to include the influencing variables, and then I just call the pairs function on the dataset.

```
#set up a subset for evaluated data
evalVar <- select(cancerData, age, number.of.sexual.partners, total.smokes, num.of.preg
nancies, stds.number)
str(evalVar)
```

```
## 'data.frame':    858 obs. of  5 variables:
##  $ age                      : int  13 14 14 14 14 14 15 15 15 15 ...
##  $ number.of.sexual.partners: int  1 2 0 1 5 1 1 1 1 1 ...
##  $ total.smokes             : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ num.of.pregnancies       : int  0 1 1 2 0 0 1 1 1 1 ...
##  $ stds.number              : int  0 0 0 0 0 0 0 0 0 0 ...
```

```
# pair correlation plot
pairs(evalVar, upper.panel = panel.cor, diag.panel = panel.hist)
```

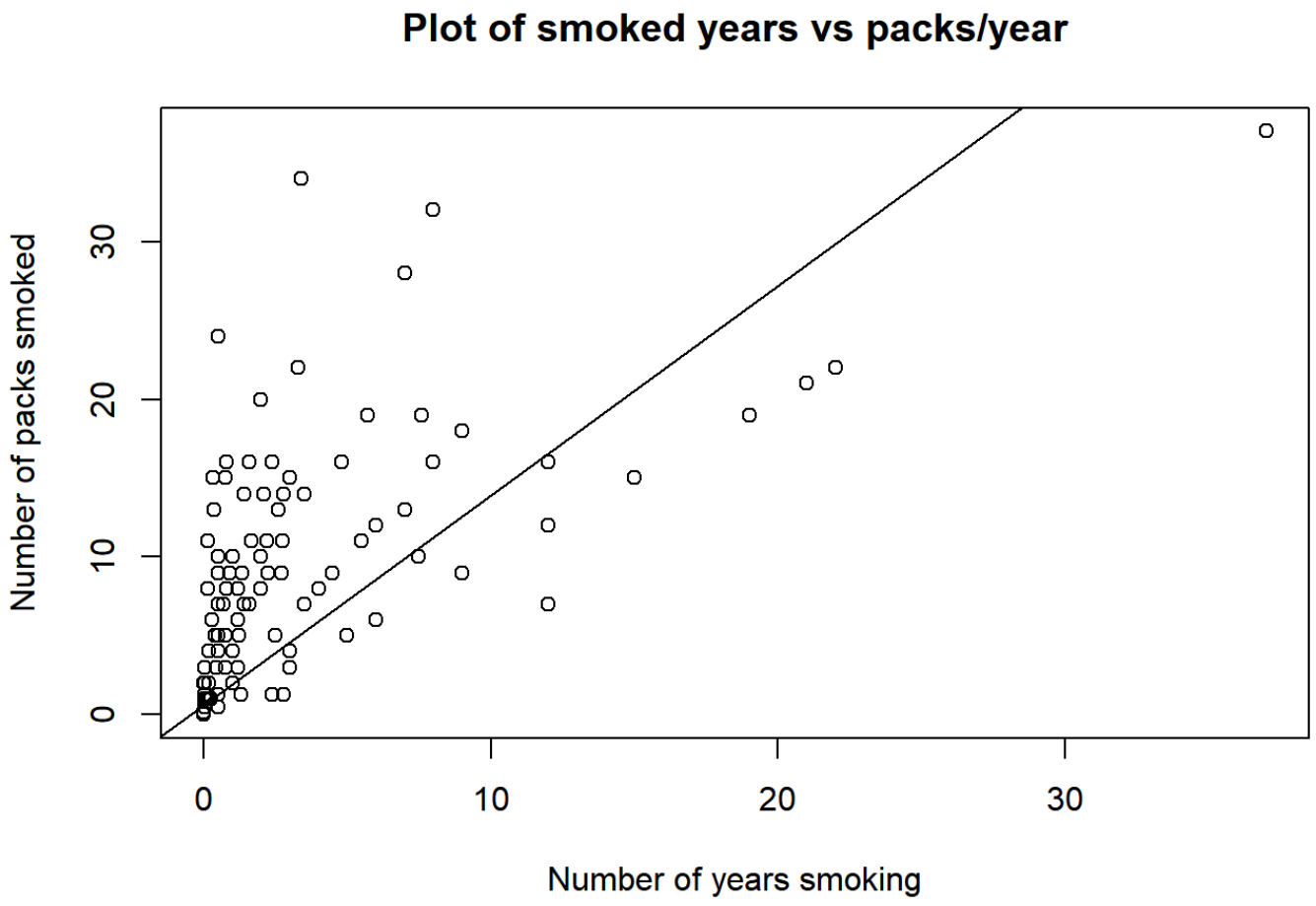


Looking at this graphic, we can tell that most influencing variables have very little correlation with each other. However, one relationship that stands out is the correlation between age and number of pregnancies, which is 0.49. One surprising thing about this graphic is the low correlation between number of sexual partners and the number of STDs the person had. I had inferred that those would have a good correlation, as having a larger number of sexual partners would put you at greater risk of catching an STD, but it appears I was incorrect.

Linear regressions

In this next section, we want to create some linear models to look at the relationship between various variables. First, I created a plot a linear regression between the number of years a person has been smoking and the number of packs they smoke per year. I then plot it and put a line of best fit on top of it. Next, I wanted to see the relationship between different forms of birth control, and whether or not people using one type would use the other. For this linear regression, used the contraceptive years and the IUD years.

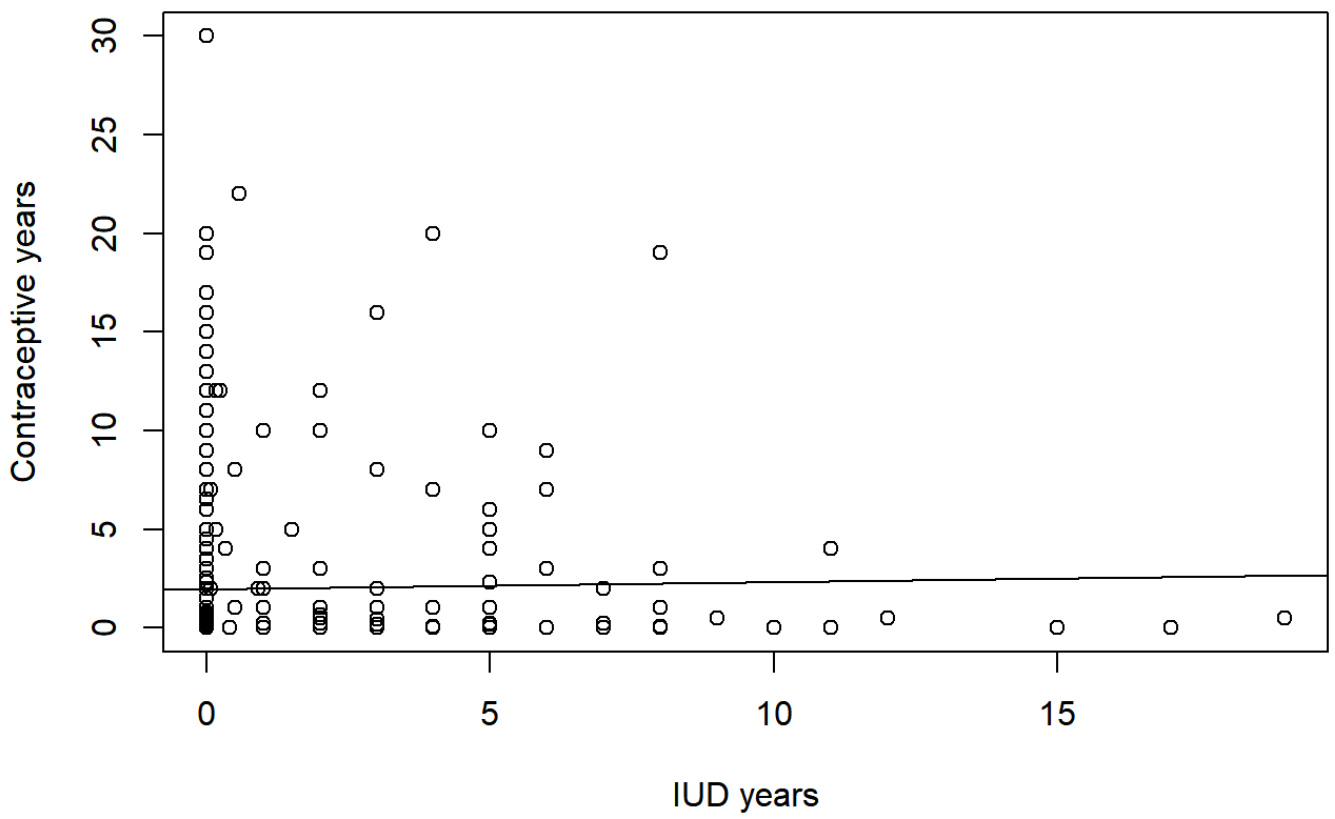
```
# linear regression model (linear regression graphing included):
#: smoke years ~ smoke packs per year
SYSP <- lm(cancerData$smokes.years~cancerData$smokes.packs.year) # creates regression subset of smoked years by smoked packs per year
plot(cancerData$smokes.years~cancerData$smokes.packs.year, main = "Plot of smoked years vs packs/year", xlab = "Number of years smoking", ylab = "Number of packs smoked") #creates a scatterplot
abline(SYSP) #creates a line of best fit off of the regression subset
```




```
summary(SYSP) #outputs stats of subset
```

```
##
## Call:
## lm(formula = cancerData$smokes.years ~ cancerData$smokes.packs.year)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.8412  -0.6074  -0.6074  -0.6074  28.8684
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    0.60740    0.09757   6.226 7.51e-10 ***
## cancerData$smokes.packs.year  1.33064    0.04329  30.737  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.801 on 856 degrees of freedom
## Multiple R-squared:  0.5246, Adjusted R-squared:  0.5241
## F-statistic: 944.7 on 1 and 856 DF,  p-value: < 2.2e-16
```

```
#: hormonal contraceptives ~ IUDs
HCIUD <- lm(cancerData$hormonal.contraceptives.years~cancerData$iud.years)
plot(cancerData$hormonal.contraceptives.years~cancerData$iud.years, main = "Plot of hor
monal contraceptive usage vs IUD usage", xlab = "IUD years", ylab = "Contraceptive year
s")
abline(HCIUD)
```




```
summary(HCIUD)
```

```
##
## Call:
## lm(formula = cancerData$hormonal.contraceptives.years ~ cancerData$iud.years)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5619 -1.9566 -1.7066  0.0434 28.0434
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           1.95656    0.12652  15.464   <2e-16 ***
## cancerData$iud.years  0.03561    0.06777   0.525    0.599
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.599 on 856 degrees of freedom
## Multiple R-squared:  0.0003224,  Adjusted R-squared:  -0.0008454
## F-statistic: 0.2761 on 1 and 856 DF,  p-value: 0.5994
```

There is a 52% correlation coefficient between the number of years someone has smoked and the number of cigarette packs they smoke per year. On the other hand, there is a <1% correlation coefficent between the different types of birth control usage, suggesting that there is not relationship between the two.

Next, I perform a mutli-regression, combining my predicted influencing variables into a linear regression model for both cervical cancer and HPV. However, I can't graph this so only the summary statistics will be outputted.

```
#: age ~ sexual partners ~ pregnancies ~ std count (multi-regression)
mC <- lm(cancerData$biopsy~evalVar$age+evalVar$number.of.sexual.partners+evalVar$total.
smokes+evalVar$num.of.pregnancies+evalVar$stds.number)
mC
```

```
##
## Call:
## lm(formula = cancerData$biopsy ~ evalVar$age + evalVar$number.of.sexual.partners +
##     evalVar$total.smokes + evalVar$num.of.pregnancies + evalVar$stds.number)
##
## Coefficients:
##                         (Intercept)                          evalVar$age
##                           1.495e-02                            1.856e-03
## evalVar$number.of.sexual.partners                evalVar$total.smokes
##                           -9.227e-04                            2.199e-05
##        evalVar$num.of.pregnancies                 evalVar$stds.number
##                           -2.808e-03                            4.803e-02
```

```
mHPV <- lm(cancerData$dx.hpv~evalVar$age+evalVar$number.of.sexual.partners+evalVar$tota
l.smokes+evalVar$num.of.pregnancies+evalVar$stds.number)
mHPV
```

```
##
## Call:
## lm(formula = cancerData$dx.hpv ~ evalVar$age + evalVar$number.of.sexual.partners +
##     evalVar$total.smokes + evalVar$num.of.pregnancies + evalVar$stds.number)
##
## Coefficients:
##                         (Intercept)                          evalVar$age
##                           -0.0175687                            0.0012301
## evalVar$number.of.sexual.partners                evalVar$total.smokes
##                            0.0008352                            0.0004176
##        evalVar$num.of.pregnancies                 evalVar$stds.number
##                            0.0005281                           -0.0035461
```

Surprisingly, the multi-regression has a very low correlation coefficient, suggesting that my predicted influencing variables do not have a large impact on the diagnosis of cervical cancer through biopsy or HPV.

BIC analysis

In this final portion of the code, I conduct a BIC, or Bayesian information criterion, analysis to look at the correlative relationship between cervical cancer, HPV, CIN, and my predicted influencing variables. I've set up the code below to print the absolute value of the difference between two BIC analysis. I run four total BIC analysis, and the specific analysis is defined in the comment above its respective code block.

```
# BIC analysis:
#: Cancer~HPV
print(abs(BIC(lm(cancerData$biopsy~1))-BIC(lm(cancerData$biopsy~cancerData$dx.hpv)))) #
conclusion: Very good relationship between the two
```

```
## [1] 15.75199
```

```
#explanation: we take the absolute value of the difference between the two BIC values

#: Cancer~CIN
print(abs(BIC(lm(cancerData$biopsy~1))-BIC(lm(cancerData$biopsy~cancerData$dx.cin)))) #
conclusion: moderate relationship between the two
```

```
## [1] 4.305622
```

```
#: Cancer ~ age + sex partners + smoke total + pregnancies
print(abs(BIC(lm(cancerData$biopsy~1))-BIC(lm(cancerData$biopsy~cancerData$age+cancerDa
ta$number.of.sexual.partners+cancerData$total.smokes+cancerData$num.of.pregnancies))))
#conclusion: very good relationship between the two
```

```
## [1] 24.1525
```

```
#: HPV ~ age + sex partners + smoke total + pregnancies
print(abs(BIC(lm(cancerData$dx.hpv~1))-BIC(lm(cancerData$dx.hpv~cancerData$age+cancerDa
ta$number.of.sexual.partners+cancerData$total.smokes+cancerData$num.of.pregnancies))))
#conclusion: moderate relationship between the two
```

```
## [1] 5.817422
```

Our first BIC analysis, between cervical cancer and HPV, returns 15.75. This is a very high number, and suggests a close relationship between the two variables. This also matches with existing reserach that suggests HPV and cervical cancer are closely linked. Our second BIC analysis, between cervical cancer and CIN (abnormal growth of cells on the surface of the cervix), scored lower - with only a 4.31 BIC

difference. While still good, this suggests a moderate relationship between existence of CIN and diagnosis of cervical cancer. For the third BIC analysis, between cervical cancer and the various selected influencing variables, our BIC differece is a whopping 24.15, which shows a nearly-causative relationship between our influencing variables and the diagnosis of cervical cancer. However, this is not reflected in our fourth (and final) BIC analysis, between HPV and our influencing variables, which is only 5.82. This is likely because cervical cancer cases are more similar to each other, while HPV is extremely wide-ranging in its effects and causes.

Extra: Mutation to append z-score columns

In this extra portion I provide the code I used to append the various z.scored numeric columns (which are at the end of the dataset). I basically just used the mutate function a bunch of types and took a Rank-z transformation of existing numeric columns but gave them new names.

```
cancerData <- mutate(cancerData, age.Z = rz.transform(cancerData$age), `first-sexual-in
tercourse.Z` = rz.transform(cancerData$`first-sexual-intercourse`), `number-of-sexual-p
artners.Z` = rz.transform(cancerData$`number-of-sexual-partners`), `num-of-pregnancies.
Z` = rz.transform(cancerData$`num-of-pregnancies`),`smokes-years.Z` = rz.transform(canc
erData$`smokes-years`),`smokes-packs-year.Z` = rz.transform(cancerData$`smokes-packs-ye
ar`), `hormonal-contraceptives-years.Z` = rz.transform(cancerData$`hormonal-contracepti
ves-years`),`iud-years.Z` = rz.transform(cancerData$`iud-years`), `stds-number.Z` = rz.
transform(cancerData$`stds-number`), `stds-number-of-diagnosis.Z` = rz.transform(cancer
Data$`stds-number-of-diagnosis`), `stds-time-since-first-diagnosis.Z` = rz.transform(ca
ncerData$`stds-time-since-first-diagnosis`), `stds-time-since-last-diagnosis.Z` = rz.tr
ansform(cancerData$`stds-time-since-last-diagnosis`))
```