Cervical_cancer_violetgo

Xinyu Wang

2023-05-03

1 Data preprocessing:

Packages

```
library(ggplot2)
library(dplyr)
```

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

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

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

```
library(Boruta)
library(DataExplorer)
library(rmarkdown)
library(flexdashboard)
library(readr)
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

Load Data

```
cervical <- read.csv("risk_factors_cervical_cancer.csv")
head(cervical,5)# first 5 rows of data
```

```
##   Age Number.of.sexual.partners First.sexual.intercourse Num.of.pregnancies
## 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
##   Smokes Smokes..years. Smokes..packs.year. Hormonal.Contraceptives
## 1    0.0            0.0                0.0                     0.0
## 2    0.0            0.0                0.0                     0.0
## 3    0.0            0.0                0.0                     0.0
## 4    1.0           37.0               37.0                     1.0
```

```
## 5    0.0            0.0                 0.0                    1.0
##   Hormonal.Contraceptives..years. IUD IUD..years. STDs STDs..number.
## 1                             0.0 0.0         0.0  0.0           0.0
## 2                             0.0 0.0         0.0  0.0           0.0
## 3                             0.0 0.0         0.0  0.0           0.0
## 4                             3.0 0.0         0.0  0.0           0.0
## 5                            15.0 0.0         0.0  0.0           0.0
##   STDs.condylomatosis STDs.cervical.condylomatosis STDs.vaginal.condylomatosis
## 1                 0.0                          0.0                        0.0
## 2                 0.0                          0.0                        0.0
## 3                 0.0                          0.0                        0.0
## 4                 0.0                          0.0                        0.0
## 5                 0.0                          0.0                        0.0
##   STDs.vulvo.perineal.condylomatosis STDs.syphilis
## 1                                0.0           0.0
## 2                                0.0           0.0
## 3                                0.0           0.0
## 4                                0.0           0.0
## 5                                0.0           0.0
##   STDs.pelvic.inflammatory.disease STDs.genital.herpes
## 1                              0.0                 0.0
## 2                              0.0                 0.0
## 3                              0.0                 0.0
## 4                              0.0                 0.0
## 5                              0.0                 0.0
##   STDs.molluscum.contagiosum STDs.AIDS STDs.HIV STDs.Hepatitis.B STDs.HPV
## 1                        0.0       0.0      0.0              0.0      0.0
## 2                        0.0       0.0      0.0              0.0      0.0
## 3                        0.0       0.0      0.0              0.0      0.0
## 4                        0.0       0.0      0.0              0.0      0.0
## 5                        0.0       0.0      0.0              0.0      0.0
##   STDs..Number.of.diagnosis STDs..Time.since.first.diagnosis
## 1                         0                               ?
## 2                         0                               ?
## 3                         0                               ?
## 4                         0                               ?
## 5                         0                               ?
##   STDs..Time.since.last.diagnosis Dx.Cancer Dx.CIN Dx.HPV Dx Hinselmann
## 1                               ?         0      0      0  0          0
## 2                               ?         0      0      0  0          0
## 3                               ?         0      0      0  0          0
## 4                               ?         1      0      1  0          0
## 5                               ?         0      0      0  0          0
##   Schiller Citology Biopsy
## 1        0        0      0
## 2        0        0      0
## 3        0        0      0
## 4        0        0      0
## 5        0        0      0
```

Columns

```
colnames(cervical) #names of variables
```

```
##  [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"
```

data info

```
glimpse(cervical)
```

```
## Rows: 858
## Columns: 36
## $ Age                                 <int> 18, 15, 34, 52, 46, 42, 51, 26, 45,~
## $ Number.of.sexual.partners           <chr> "4.0", "1.0", "1.0", "5.0", "3.0", ~
## $ First.sexual.intercourse            <chr> "15.0", "14.0", "?", "16.0", "21.0"~
## $ Num.of.pregnancies                  <chr> "1.0", "1.0", "1.0", "4.0", "4.0", ~
## $ Smokes                              <chr> "0.0", "0.0", "0.0", "1.0", "0.0", ~
## $ Smokes..years.                      <chr> "0.0", "0.0", "0.0", "37.0", "0.0",~
## $ Smokes..packs.year.                 <chr> "0.0", "0.0", "0.0", "37.0", "0.0",~
## $ Hormonal.Contraceptives             <chr> "0.0", "0.0", "0.0", "1.0", "1.0", ~
## $ Hormonal.Contraceptives..years.     <chr> "0.0", "0.0", "0.0", "3.0", "15.0",~
## $ IUD                                 <chr> "0.0", "0.0", "0.0", "0.0", "0.0", ~
## $ IUD..years.                         <chr> "0.0", "0.0", "0.0", "0.0", "0.0", ~
## $ STDs                                <chr> "0.0", "0.0", "0.0", "0.0", "0.0", ~
## $ STDs..number.                       <chr> "0.0", "0.0", "0.0", "0.0", "0.0", ~
## $ STDs.condylomatosis                 <chr> "0.0", "0.0", "0.0", "0.0", "0.0", ~
## $ STDs.cervical.condylomatosis        <chr> "0.0", "0.0", "0.0", "0.0", "0.0", ~
## $ STDs.vaginal.condylomatosis         <chr> "0.0", "0.0", "0.0", "0.0", "0.0", ~
## $ STDs.vulvo.perineal.condylomatosis <chr> "0.0", "0.0", "0.0", "0.0", "0.0", ~
## $ STDs.syphilis                       <chr> "0.0", "0.0", "0.0", "0.0", "0.0", ~
## $ STDs.pelvic.inflammatory.disease    <chr> "0.0", "0.0", "0.0", "0.0", "0.0", ~
## $ STDs.genital.herpes                 <chr> "0.0", "0.0", "0.0", "0.0", "0.0", ~
## $ STDs.molluscum.contagiosum          <chr> "0.0", "0.0", "0.0", "0.0", "0.0", ~
## $ STDs.AIDS                           <chr> "0.0", "0.0", "0.0", "0.0", "0.0", ~
## $ STDs.HIV                            <chr> "0.0", "0.0", "0.0", "0.0", "0.0", ~
## $ STDs.Hepatitis.B                    <chr> "0.0", "0.0", "0.0", "0.0", "0.0", ~
## $ STDs.HPV                            <chr> "0.0", "0.0", "0.0", "0.0", "0.0", ~
## $ STDs..Number.of.diagnosis           <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ STDs..Time.since.first.diagnosis    <chr> "?", "?", "?", "?", "?", "?", "?", ~
## $ STDs..Time.since.last.diagnosis     <chr> "?", "?", "?", "?", "?", "?", "?", ~
## $ Dx.Cancer                           <int> 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0,~
```

```
## $ Dx.CIN                            <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ Dx.HPV                            <int> 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0,~
## $ Dx                                <int> 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0,~
## $ Hinselmann                        <int> 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0,~
## $ Schiller                          <int> 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0,~
## $ Citology                          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ Biopsy                            <int> 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0,~
```

```
summary(cervical)
```

```
##       Age        Number.of.sexual.partners First.sexual.intercourse
##  Min.   :13.00   Length:858                Length:858
##  1st Qu.:20.00   Class :character          Class :character
##  Median :25.00   Mode  :character          Mode  :character
##  Mean   :26.82
##  3rd Qu.:32.00
##  Max.   :84.00
##  Num.of.pregnancies    Smokes          Smokes..years.     Smokes..packs.year.
##  Length:858         Length:858         Length:858         Length:858
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##  Hormonal.Contraceptives Hormonal.Contraceptives..years.     IUD
##  Length:858              Length:858                      Length:858
##  Class :character        Class :character                Class :character
##  Mode  :character        Mode  :character                Mode  :character
##
##
##
##  IUD..years.            STDs           STDs..number.      STDs.condylomatosis
##  Length:858         Length:858         Length:858         Length:858
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##  STDs.cervical.condylomatosis STDs.vaginal.condylomatosis
##  Length:858                   Length:858
##  Class :character             Class :character
##  Mode  :character             Mode  :character
##
##
##
##  STDs.vulvo.perineal.condylomatosis STDs.syphilis
##  Length:858                         Length:858
##  Class :character                   Class :character
##  Mode  :character                   Mode  :character
##
##
##
##  STDs.pelvic.inflammatory.disease STDs.genital.herpes
##  Length:858                       Length:858
##  Class :character                 Class :character
```

```
##  Mode  :character                 Mode  :character
##
##
##
##  STDs.molluscum.contagiosum  STDs.AIDS           STDs.HIV
##  Length:858                  Length:858         Length:858
##  Class :character            Class :character   Class :character
##  Mode  :character            Mode  :character   Mode  :character
##
##
##
##  STDs.Hepatitis.B     STDs.HPV         STDs..Number.of.diagnosis
##  Length:858         Length:858         Min.   :0.00000
##  Class :character   Class :character   1st Qu.:0.00000
##  Mode  :character   Mode  :character   Median :0.00000
##                                        Mean   :0.08741
##                                        3rd Qu.:0.00000
##                                        Max.   :3.00000
##  STDs..Time.since.first.diagnosis STDs..Time.since.last.diagnosis
##  Length:858                       Length:858
##  Class :character                 Class :character
##  Mode  :character                 Mode  :character
##
##
##
##    Dx.Cancer           Dx.CIN            Dx.HPV              Dx
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000
##  Median :0.00000   Median :0.00000   Median :0.00000   Median :0.00000
##  Mean   :0.02098   Mean   :0.01049   Mean   :0.02098   Mean   :0.02797
##  3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.00000   Max.   :1.00000
##    Hinselmann         Schiller          Citology           Biopsy
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.0000
##  1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.0000
##  Median :0.00000   Median :0.00000   Median :0.00000   Median :0.0000
##  Mean   :0.04079   Mean   :0.08625   Mean   :0.05128   Mean   :0.0641
##  3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.0000
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.00000   Max.   :1.0000
```

delete characters like "?" since this can lead to char

```
for(i in 1:ncol(cervical)) {
  for (j in 1:nrow(cervical)){
  if (cervical[j,i] == "?"){
  cervical[j , i] <- NA}
  else next
  }}
head(cervical)
```

```
##   Age Number.of.sexual.partners First.sexual.intercourse Num.of.pregnancies
## 1  18                       4.0                     15.0                1.0
## 2  15                       1.0                     14.0                1.0
## 3  34                       1.0                     <NA>                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
##   Smokes Smokes..years. Smokes..packs.year. Hormonal.Contraceptives
## 1    0.0           0.0                0.0                     0.0
## 2    0.0           0.0                0.0                     0.0
## 3    0.0           0.0                0.0                     0.0
## 4    1.0          37.0               37.0                     1.0
## 5    0.0           0.0                0.0                     1.0
## 6    0.0           0.0                0.0                     0.0
##   Hormonal.Contraceptives..years. IUD IUD..years. STDs STDs..number.
## 1                             0.0 0.0         0.0  0.0           0.0
## 2                             0.0 0.0         0.0  0.0           0.0
## 3                             0.0 0.0         0.0  0.0           0.0
## 4                             3.0 0.0         0.0  0.0           0.0
## 5                            15.0 0.0         0.0  0.0           0.0
## 6                             0.0 0.0         0.0  0.0           0.0
##   STDs.condylomatosis STDs.cervical.condylomatosis STDs.vaginal.condylomatosis
## 1                 0.0                          0.0                        0.0
## 2                 0.0                          0.0                        0.0
## 3                 0.0                          0.0                        0.0
## 4                 0.0                          0.0                        0.0
## 5                 0.0                          0.0                        0.0
## 6                 0.0                          0.0                        0.0
##   STDs.vulvo.perineal.condylomatosis STDs.syphilis
## 1                                0.0           0.0
## 2                                0.0           0.0
## 3                                0.0           0.0
## 4                                0.0           0.0
## 5                                0.0           0.0
## 6                                0.0           0.0
##   STDs.pelvic.inflammatory.disease STDs.genital.herpes
## 1                              0.0                 0.0
## 2                              0.0                 0.0
## 3                              0.0                 0.0
## 4                              0.0                 0.0
## 5                              0.0                 0.0
## 6                              0.0                 0.0
##   STDs.molluscum.contagiosum STDs.AIDS STDs.HIV STDs.Hepatitis.B STDs.HPV
## 1                        0.0       0.0      0.0              0.0      0.0
## 2                        0.0       0.0      0.0              0.0      0.0
## 3                        0.0       0.0      0.0              0.0      0.0
## 4                        0.0       0.0      0.0              0.0      0.0
## 5                        0.0       0.0      0.0              0.0      0.0
## 6                        0.0       0.0      0.0              0.0      0.0
##   STDs..Number.of.diagnosis STDs..Time.since.first.diagnosis
## 1                         0                             <NA>
## 2                         0                             <NA>
## 3                         0                             <NA>
## 4                         0                             <NA>
## 5                         0                             <NA>
## 6                         0                             <NA>
##   STDs..Time.since.last.diagnosis Dx.Cancer Dx.CIN Dx.HPV Dx Hinselmann
## 1                            <NA>         0      0      0  0          0
```

```
## 2                           <NA>         0      0      0  0          0
## 3                           <NA>         0      0      0  0          0
## 4                           <NA>         1      0      1  0          0
## 5                           <NA>         0      0      0  0          0
## 6                           <NA>         0      0      0  0          0
##   Schiller Citology Biopsy
## 1        0        0      0
## 2        0        0      0
## 3        0        0      0
## 4        0        0      0
## 5        0        0      0
## 6        0        0      0
```

to numeric

```
# ro numeric
cervical <- apply(cervical,2,as.numeric)
cervical <- round(cervical,digits = 2)
cervical <- as.data.frame(cervical)
# some cols to factors
cervical$Smokes <- as.factor(cervical$Smokes)
cervical$Hormonal.Contraceptives <- as.factor(cervical$Hormonal.Contraceptives)
cervical$IUD <- as.factor(cervical$IUD)

cervical$STDs <- as.factor(cervical$STDs)
cervical$STDs.condylomatosis <- as.factor(cervical$STDs.condylomatosis)
cervical$STDs.vaginal.condylomatosis <- as.factor(cervical$STDs.vaginal.condylomatosis)
cervical$STDs.vulvo.perineal.condylomatosis <- as.factor(cervical$STDs.condylomatosis)
cervical$STDs.condylomatosis <- as.factor(cervical$STDs.vulvo.perineal.condylomatosis)
cervical$STDs.syphilis  <- as.factor(cervical$STDs.syphilis)
cervical$STDs.pelvic.inflammatory.disease <- as.factor(cervical$STDs.pelvic.inflammatory.disease)
cervical$STDs.genital.herpes <- as.factor(cervical$STDs.genital.herpes)
cervical$STDs.molluscum.contagiosum <- as.factor(cervical$STDs.molluscum.contagiosum)
cervical$STDs.AIDS <- as.factor(cervical$STDs.AIDS)
cervical$STDs.HIV <- as.factor(cervical$STDs.HIV)
cervical$STDs.Hepatitis.B <- as.factor(cervical$STDs.Hepatitis.B)
cervical$STDs.HPV <- as.factor(cervical$STDs.HPV)
# Dx changed as bool
cervical$Dx.Cancer <- as.factor(cervical$Dx.Cancer)
cervical$Dx.CIN <- as.factor(cervical$Dx.CIN)
cervical$Dx.HPV <- as.factor(cervical$Dx)
cervical$Hinselmann <- as.factor(cervical$Hinselmann)
cervical$Schiller <- as.factor(cervical$Schiller)
cervical$Citology <- as.factor(cervical$Citology)
cervical$Biopsy <- as.factor(cervical$Biopsy)
head(cervical,5)
```

```
##   Age Number.of.sexual.partners First.sexual.intercourse Num.of.pregnancies
## 1  18                         4                       15                   1
## 2  15                         1                       14                   1
## 3  34                         1                       NA                   1
## 4  52                         5                       16                   4
## 5  46                         3                       21                   4
##   Smokes Smokes..years. Smokes..packs.year. Hormonal.Contraceptives
```

```
## 1      0              0                  0                      0
## 2      0              0                  0                      0
## 3      0              0                  0                      0
## 4      1             37                 37                      1
## 5      0              0                  0                      1
##   Hormonal.Contraceptives..years. IUD IUD..years. STDs STDs..number.
## 1                               0   0           0    0             0
## 2                               0   0           0    0             0
## 3                               0   0           0    0             0
## 4                               3   0           0    0             0
## 5                              15   0           0    0             0
##   STDs.condylomatosis STDs.cervical.condylomatosis STDs.vaginal.condylomatosis
## 1                   0                            0                          0
## 2                   0                            0                          0
## 3                   0                            0                          0
## 4                   0                            0                          0
## 5                   0                            0                          0
##   STDs.vulvo.perineal.condylomatosis STDs.syphilis
## 1                                  0             0
## 2                                  0             0
## 3                                  0             0
## 4                                  0             0
## 5                                  0             0
##   STDs.pelvic.inflammatory.disease STDs.genital.herpes
## 1                                0                   0
## 2                                0                   0
## 3                                0                   0
## 4                                0                   0
## 5                                0                   0
##   STDs.molluscum.contagiosum STDs.AIDS STDs.HIV STDs.Hepatitis.B STDs.HPV
## 1                          0         0        0                0        0
## 2                          0         0        0                0        0
## 3                          0         0        0                0        0
## 4                          0         0        0                0        0
## 5                          0         0        0                0        0
##   STDs..Number.of.diagnosis STDs..Time.since.first.diagnosis
## 1                         0                              NA
## 2                         0                              NA
## 3                         0                              NA
## 4                         0                              NA
## 5                         0                              NA
##   STDs..Time.since.last.diagnosis Dx.Cancer Dx.CIN Dx.HPV Dx Hinselmann
## 1                              NA         0      0      0  0          0
## 2                              NA         0      0      0  0          0
## 3                              NA         0      0      0  0          0
## 4                              NA         1      0      0  0          0
## 5                              NA         0      0      0  0          0
##   Schiller Citology Biopsy
## 1        0        0      0
## 2        0        0      0
## 3        0        0      0
## 4        0        0      0
## 5        0        0      0
```

missing values

```
# delete if too much missing values and leave it as it is if not
na_prop <- colMeans(is.na(cervical))
cervical <- cervical[,na_prop <= 0.8]
cervical_storage <- cervical
cervical_withNA <- cervical
cervical_withMean <- cervical
```

```
#filling the value NA
for (col_name in names(cervical_withMean)) {
  # if it
  if (is.numeric(cervical_withMean[[col_name]])) {
    mean_value <- mean(cervical_withMean[[col_name]], na.rm = TRUE)
    cervical_withMean[[col_name]][is.na(cervical_withMean[[col_name]])] <- mean_value
  }
  # if col is factor, then use random sampling
  else if (is.factor(cervical_withMean[[col_name]])) {
    set.seed(42)
    cervical_withMean[[col_name]][is.na(cervical_withMean[[col_name]])] <- sample(cervical_withMean[[co
                                                          sum(is.na(cervical_withMean[[col_name]])),
                                                          replace = TRUE)
  }
}
```

```
summary(cervical_withMean)
```

```
##       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     0:734   Min.   : 0.00   Min.   : 0.0000
##  1st Qu.: 1.000     1:124   1st Qu.: 0.00   1st Qu.: 0.0000
##  Median : 2.000             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
##  0:313                   Min.   : 0.000                  0:765
##  1:545                   1st Qu.: 0.000                  1: 93
##                          Median : 1.000
##                          Mean   : 2.256
##                          3rd Qu.: 2.256
##                          Max.   :30.000
##   IUD..years.      STDs    STDs..number.    STDs.condylomatosis
##  Min.   : 0.0000   0:769   Min.   :0.0000   0:811
##  1st Qu.: 0.0000   1: 89   1st Qu.:0.0000   1: 47
##  Median : 0.0000           Median :0.0000
##  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
##  Min.   :0                    0:853
##  1st Qu.:0                    1:  5
##  Median :0
##  Mean   :0
##  3rd Qu.:0
##  Max.   :0
##  STDs.vulvo.perineal.condylomatosis STDs.syphilis
##  0:811                              0:837
##  1: 47                              1: 21
##
##
##
##
##  STDs.pelvic.inflammatory.disease STDs.genital.herpes
##  0:857                            0:857
##  1:  1                            1:  1
##
##
##
##
##  STDs.molluscum.contagiosum STDs.AIDS STDs.HIV STDs.Hepatitis.B STDs.HPV
##  0:857                      0:858     0:837    0:857            0:855
##  1:  1                                1: 21    1:  1            1:  3
##
##
##
##
##  STDs..Number.of.diagnosis Dx.Cancer Dx.CIN  Dx.HPV        Dx
##  Min.   :0.00000           0:840     0:849   0:834   Min.   :0.00000
##  1st Qu.:0.00000           1: 18     1:  9   1: 24   1st Qu.:0.00000
##  Median :0.00000                                     Median :0.00000
##  Mean   :0.08741                                     Mean   :0.02797
##  3rd Qu.:0.00000                                     3rd Qu.:0.00000
##  Max.   :3.00000                                     Max.   :1.00000
##  Hinselmann Schiller Citology Biopsy
##  0:823      0:784    0:814    0:803
##  1: 35      1: 74    1: 44    1: 55
##
##
##
##
```

delete columns with all zeros

standardization

```
# Remove columns with all values equal to zero before standardization
cervical_without_zero_cols <- cervical_withMean[, !colSums(cervical_withMean == 0, na.rm = TRUE) == nrow

# Standardize the numeric columns
for (col_name in names(cervical_without_zero_cols)) {
  if (is.numeric(cervical_without_zero_cols[[col_name]])) {
    min_value <- min(cervical_without_zero_cols[[col_name]], na.rm = TRUE)
```

```
    max_value <- max(cervical_without_zero_cols[[col_name]], na.rm = TRUE)
    cervical_without_zero_cols[[col_name]] <- (cervical_without_zero_cols[[col_name]] - min_value) / (ma
  }
}

# Update the dataset
cervical_std <- cervical_without_zero_cols

# Display the first few rows of the standardized dataset
head(cervical_std)
```

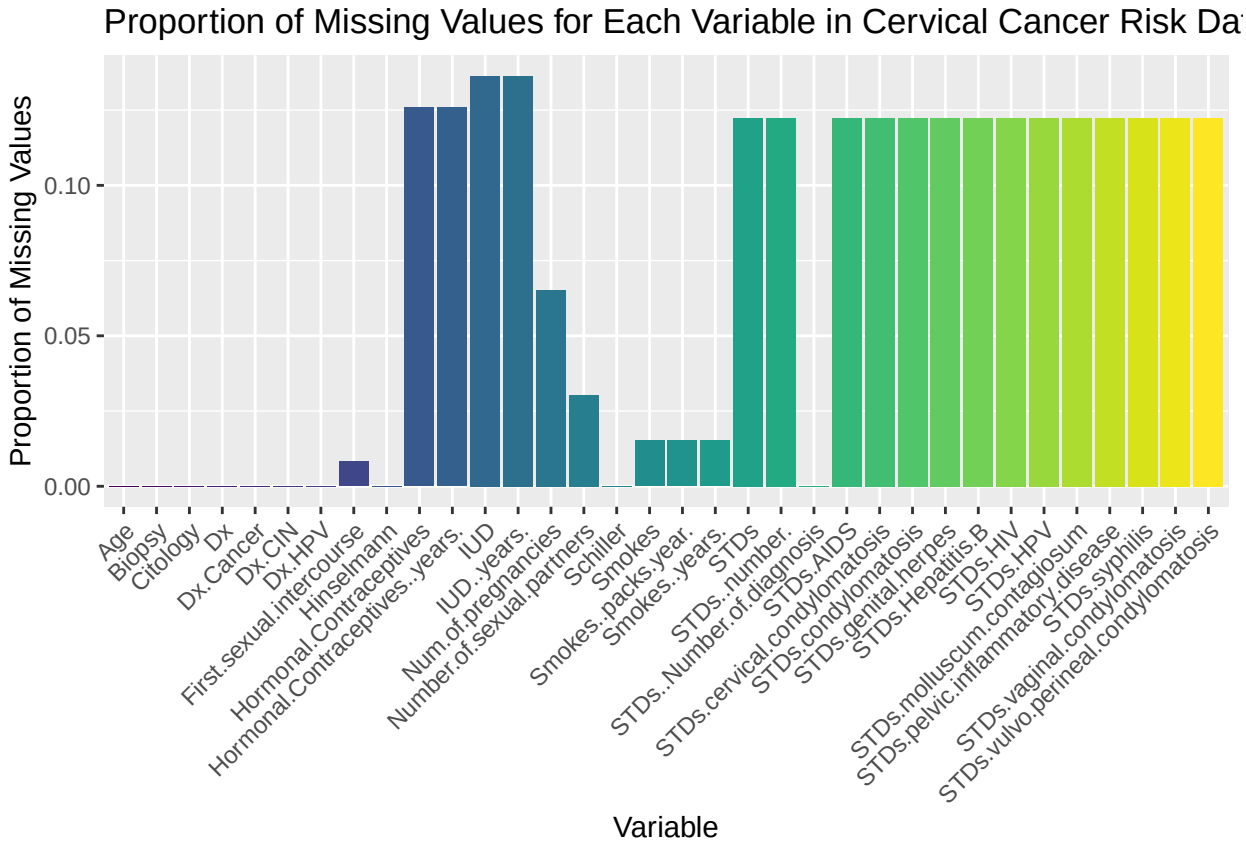
```
##          Age Number.of.sexual.partners First.sexual.intercourse
## 1 0.07042254                0.11111111                0.2272727
## 2 0.02816901                0.00000000                0.1818182
## 3 0.29577465                0.00000000                0.3179682
## 4 0.54929577                0.14814815                0.2727273
## 5 0.46478873                0.07407407                0.5000000
## 6 0.40845070                0.07407407                0.5909091
##   Num.of.pregnancies Smokes Smokes..years. Smokes..packs.year.
## 1         0.09090909      0              0                   0
## 2         0.09090909      0              0                   0
## 3         0.09090909      0              0                   0
## 4         0.36363636      1              1                   1
## 5         0.36363636      0              0                   0
## 6         0.18181818      0              0                   0
##   Hormonal.Contraceptives Hormonal.Contraceptives..years. IUD IUD..years. STDs
## 1                       0                             0.0   0           0    0
## 2                       0                             0.0   0           0    0
## 3                       0                             0.0   0           0    0
## 4                       1                             0.1   0           0    0
## 5                       1                             0.5   0           0    0
## 6                       0                             0.0   0           0    0
##   STDs..number. STDs.condylomatosis STDs.vaginal.condylomatosis
## 1             0                   0                           0
## 2             0                   0                           0
## 3             0                   0                           0
## 4             0                   0                           0
## 5             0                   0                           0
## 6             0                   0                           0
##   STDs.vulvo.perineal.condylomatosis STDs.syphilis
## 1                                  0             0
## 2                                  0             0
## 3                                  0             0
## 4                                  0             0
## 5                                  0             0
## 6                                  0             0
##   STDs.pelvic.inflammatory.disease STDs.genital.herpes
## 1                                0                   0
## 2                                0                   0
## 3                                0                   0
## 4                                0                   0
## 5                                0                   0
## 6                                0                   0
##   STDs.molluscum.contagiosum STDs.HIV STDs.Hepatitis.B STDs.HPV
```

```
## 1                           0        0               0        0
## 2                           0        0               0        0
## 3                           0        0               0        0
## 4                           0        0               0        0
## 5                           0        0               0        0
## 6                           0        0               0        0
##   STDs..Number.of.diagnosis Dx.Cancer Dx.CIN Dx.HPV Dx Hinselmann Schiller
## 1                         0         0      0      0  0          0        0
## 2                         0         0      0      0  0          0        0
## 3                         0         0      0      0  0          0        0
## 4                         0         1      0      0  0          0        0
## 5                         0         0      0      0  0          0        0
## 6                         0         0      0      0  0          0        0
##   Citology Biopsy
## 1        0      0
## 2        0      0
## 3        0      0
## 4        0      0
## 5        0      0
## 6        0      0
```

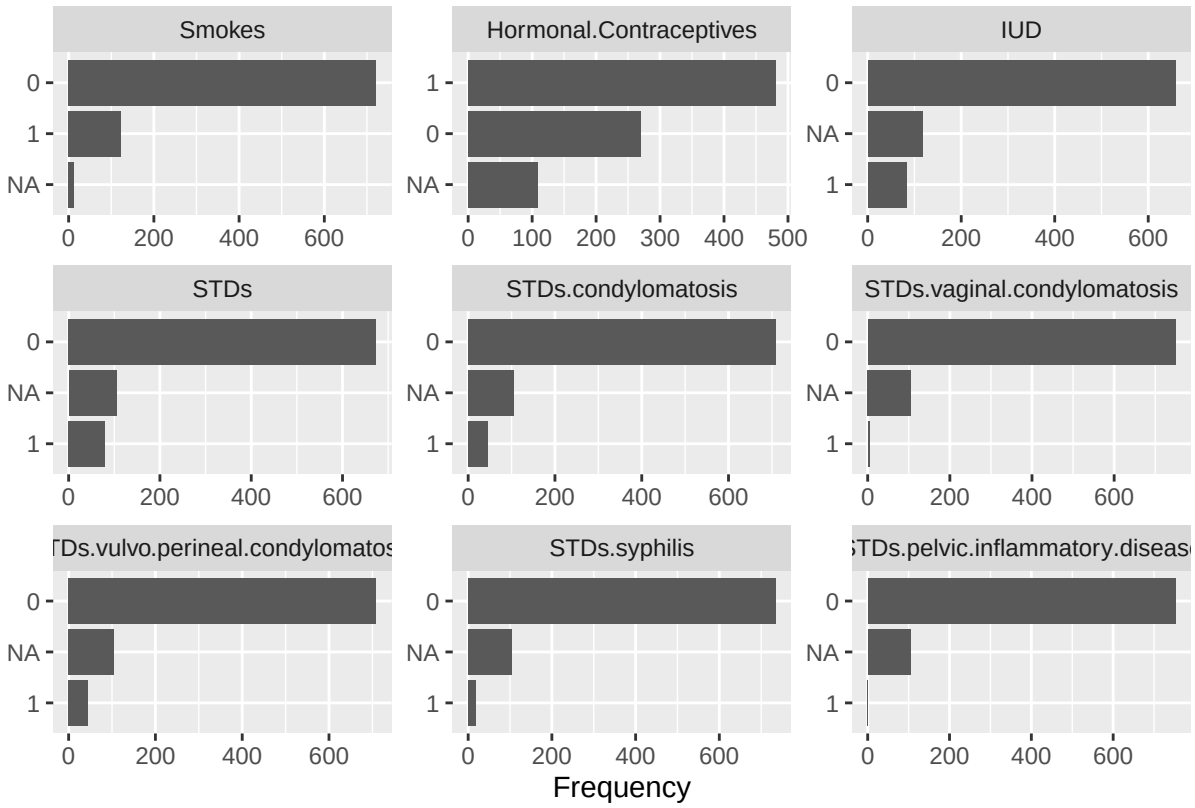
2 EDA

```
library(ggtext)

prop_NA <- function(x) { mean(is.na(x)) }
missdata <- sapply(cervical, prop_NA)
missdata <- data.frame(variable = names(missdata), prop_NA = missdata)
missdata <- missdata[order(missdata$prop_NA, decreasing = TRUE), ]

ggplot(missdata, aes(x = variable, y = prop_NA, fill = variable)) +
  geom_bar(stat = "identity", position = position_dodge(width = 0.9)) +
  scale_fill_viridis_d(guide = FALSE) +
  labs(x = "Variable", y = "Proportion of Missing Values",
       title = "Proportion of Missing Values for Each Variable in Cervical Cancer Risk Dataset") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

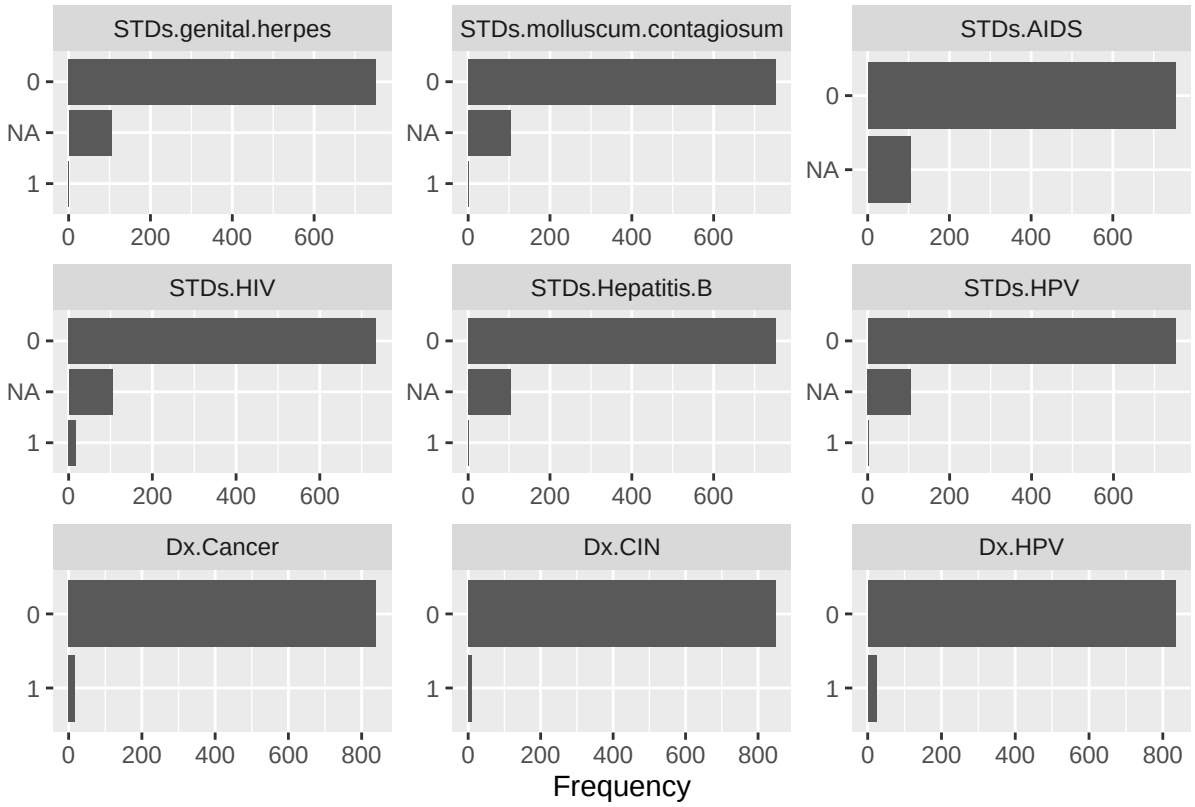
```
## Warning: The `guide` argument in `scale_*()` cannot be `FALSE`. This was deprecated in
## ggplot2 3.3.4.
## i Please use "none" instead.
```



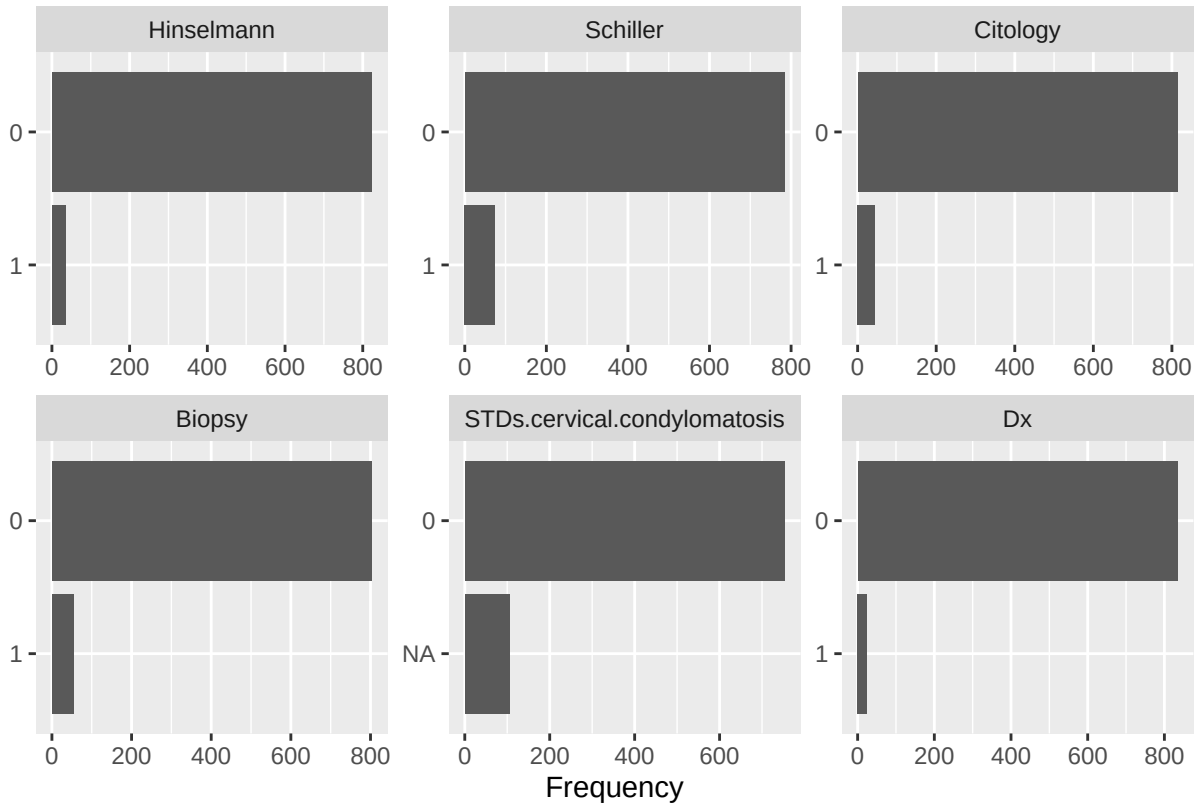

```
plot_str(cervical)
plot_bar(cervical)
```


Smokes
Hormonal.Contraceptives
IUD
STDs
STDs.condylomatosis
STDs.vaginal.condylomatosis
TDs.vulvo.perineal.condylomatos
STDs.syphilis
STDs.pelvic.inflammatory.diseas
Frequency

Page 1


STDs.genital.herpes
STDs.molluscum.contagiosum
STDs.AIDS
STDs.HIV
STDs.Hepatitis.B
STDs.HPV
Dx.Cancer
Dx.CIN
Dx.HPV
Frequency

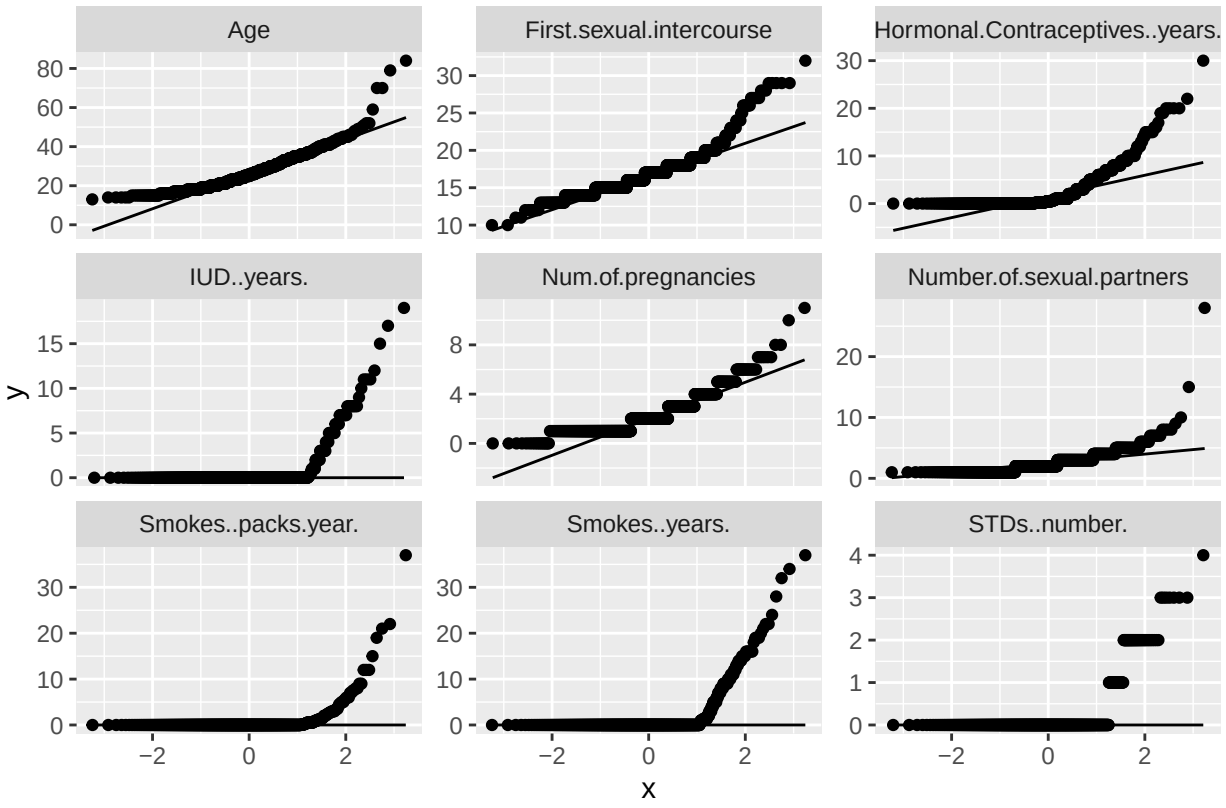
Page 2




```
plot_qq(cervical)
```

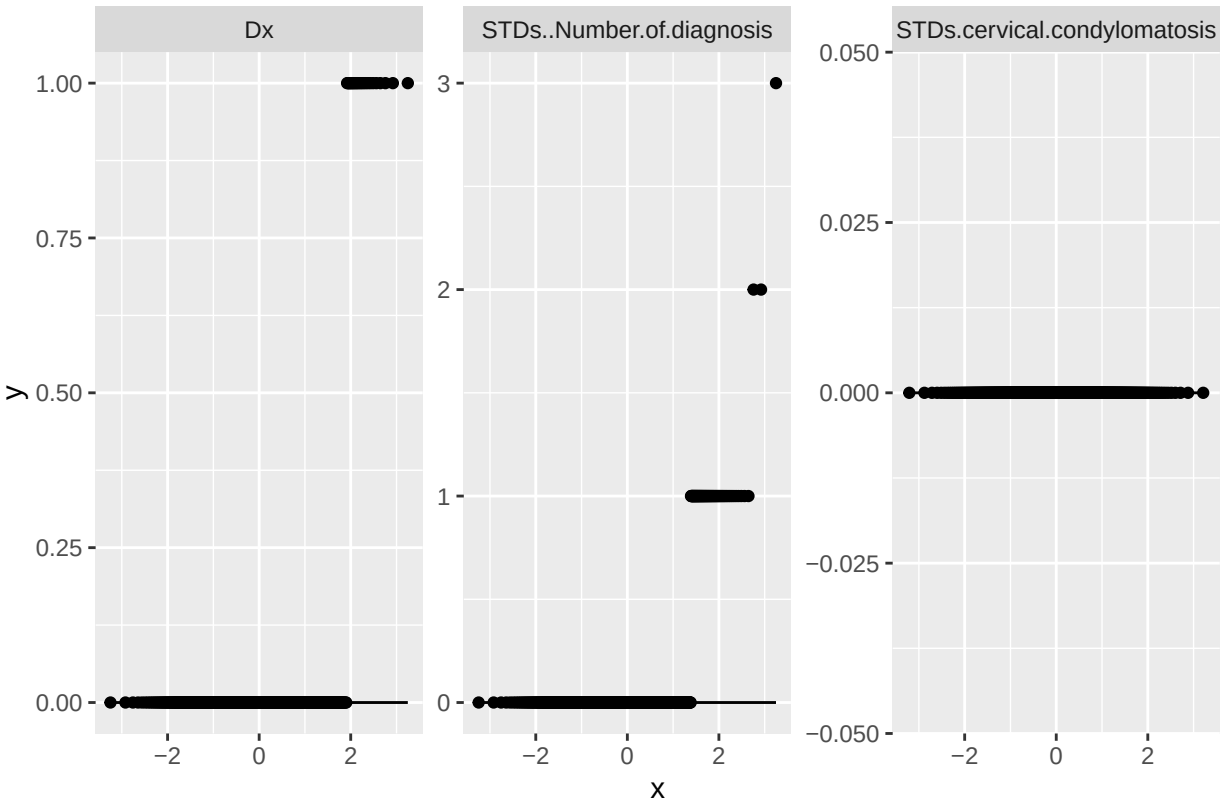
```
## Warning: Removed 445 rows containing non-finite values (`stat_qq()`).
```

```
## Warning: Removed 445 rows containing non-finite values (`stat_qq_line()`).
```



Page 1

```
## Warning: Removed 105 rows containing non-finite values (`stat_qq()`).
## Warning: Removed 105 rows containing non-finite values (`stat_qq_line()`).
```



3 Feature Selection

Define the target response

```
Cervical <- cervical_std
Dx_dt <- Cervical[,c("Hinselmann","Schiller","Citology","Biopsy")]
Cervical$Outcome <- apply(Dx_dt, 1, FUN = function(x) {return(sum(as.numeric(x)))})
```

Proportion of the Outcome with not 0 and 0

```
# Convert the "Outcome" column to numeric data type
Cervical$Outcome <- as.numeric(Cervical$Outcome)
Cervical$Outcome <- as.numeric(Cervical$Outcome)

prop <- nrow(Cervical[Cervical$Outcome==0,])/nrow(Cervical[Cervical$Outcome>0,])
cat("The ratio between control and case is", prop, "for the whole data suppose case when outcome is larg
```

```
## The ratio between control and case is 7.411765 for the whole data suppose case when outcome is larger
```

Split into training and testing set

```
Npop <- nrow(Cervical)
test_ind <- sample(Npop, Npop/5)
cervical_te <- Cervical[test_ind,]
cervical_tr <- Cervical[-test_ind,]
```

Downsampling for training set and test set(feature selection)

Downsampling for traing set

```
# Load the necessary library
library(dplyr)

# Separate cases and controls
cases <- cervical_tr[cervical_tr$Outcome != 0,]
controls <- cervical_tr[cervical_tr$Outcome == 0,]

# Function to downsample controls for different ratios
downsample_controls <- function(ratio) {
  n_controls <- nrow(cases) * ratio
  return(sample_n(controls, n_controls))
}

# Perform downsampling
controls_1_2 <- downsample_controls(2)
controls_1_3 <- downsample_controls(3)
controls_1_4 <- downsample_controls(4)
controls_1_5 <- downsample_controls(5)

# Combine cases and downsampled controls for each ratio
cervical_tr_1_2 <- rbind(cases, controls_1_2)
cervical_tr_1_3 <- rbind(cases, controls_1_3)
cervical_tr_1_4 <- rbind(cases, controls_1_4)
cervical_tr_1_5 <- rbind(cases, controls_1_5)

# Optional: Shuffle the rows (if needed)
cervical_tr_1_2 <- cervical_tr_1_2[sample(nrow(cervical_tr_1_2)),]
cervical_tr_1_3 <- cervical_tr_1_3[sample(nrow(cervical_tr_1_3)),]
cervical_tr_1_4 <- cervical_tr_1_4[sample(nrow(cervical_tr_1_4)),]
cervical_tr_1_5 <- cervical_tr_1_5[sample(nrow(cervical_tr_1_5)),]
#We choose cervical_tr_1_3 for model(feature) selection.
```

##Downsampling for test set

```
# Separate cases and controls
cases_te <- cervical_te[cervical_tr$Outcome != 0,]
controls_te <- cervical_te[cervical_tr$Outcome == 0,]

# Function to downsample controls for different ratios
downsample_controls <- function(ratio) {
  n_controls <- nrow(cases) * ratio
  return(sample_n(controls, n_controls))
}

# Perform downsampling
controls_1_2_te <- downsample_controls(2)
controls_1_3_te <- downsample_controls(3)
controls_1_4_te <- downsample_controls(4)
controls_1_5_te <- downsample_controls(5)

# Combine cases and downsampled controls for each ratio
```

```
cervical_te_1_2 <- rbind(cases, controls_1_2)
cervical_te_1_3 <- rbind(cases, controls_1_3)
cervical_te_1_4 <- rbind(cases, controls_1_4)
cervical_te_1_5 <- rbind(cases, controls_1_5)

# Optional: Shuffle the rows (if needed)
cervical_te_1_2 <- cervical_te_1_2[sample(nrow(cervical_tr_1_2)),]
cervical_te_1_3 <- cervical_te_1_3[sample(nrow(cervical_tr_1_3)),]
cervical_te_1_4 <- cervical_te_1_4[sample(nrow(cervical_tr_1_4)),]
cervical_te_1_5 <- cervical_te_1_5[sample(nrow(cervical_tr_1_5)),]
#We choose cervical_tr_1_3 for model(feature) selection.
```

Delete the columns with all zeros in cervical_tr_1_3

```
# Remove columns with all zeros
cervical_tr_1_2 <- cervical_tr_1_2[, colSums(cervical_tr_1_3 != 0) > 0]
cervical_tr_1_3 <- cervical_tr_1_3[, colSums(cervical_tr_1_3 != 0) > 0]
cervical_tr_1_4 <- cervical_tr_1_4[, colSums(cervical_tr_1_3 != 0) > 0]
cervical_tr_1_5 <- cervical_tr_1_5[, colSums(cervical_tr_1_3 != 0) > 0]
```

LASSO feature selection

```
X_tr <- cervical_tr_1_4[,1:24]
Y_tr <- cervical_tr_1_4["Outcome"]
X_te <- cervical_te_1_4[,1:24]
Y_te <- cervical_te_1_4["Outcome"]
X_class <- data.frame(sapply(X_tr, class))
```

```
library(glmnet)
```

```
## Loading required package: Matrix
```

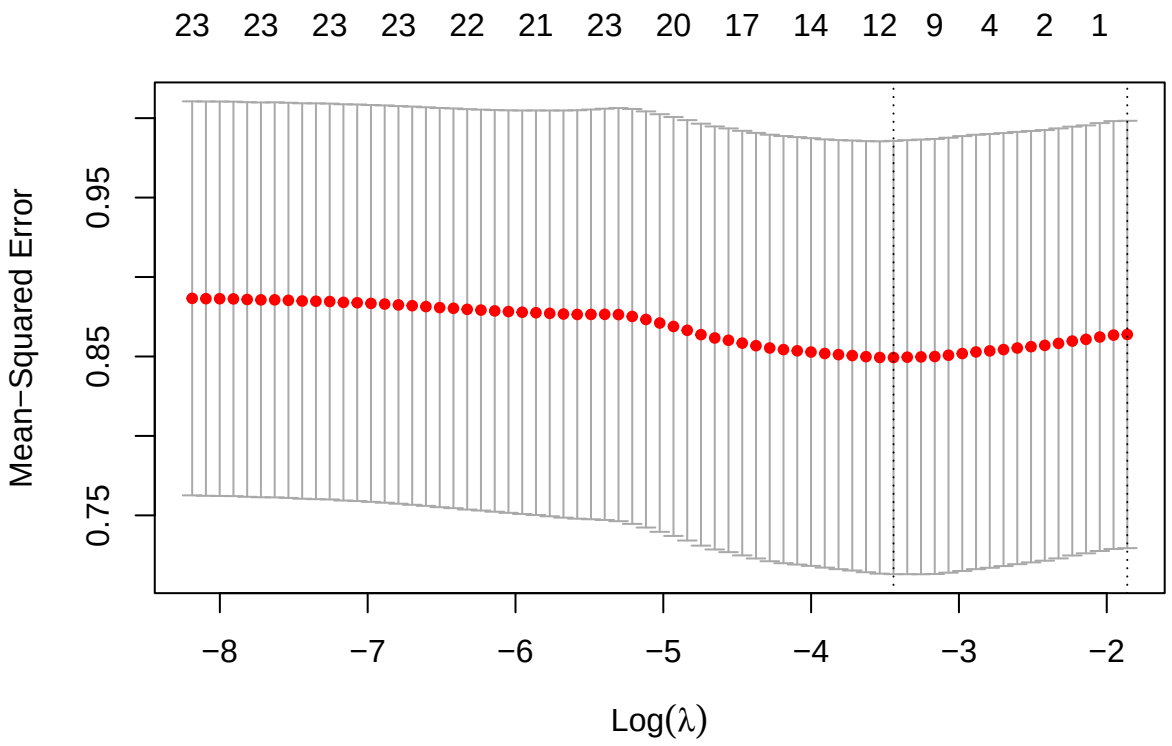
```
## Loaded glmnet 4.1-7
```

```
# input X should be a matrix
numeric_df <- X_tr[, sapply(X_tr, is.numeric)]
numeric_mat <- as.matrix(numeric_df)
factor_df <- sapply(X_tr[, sapply(X_tr, is.factor)], as.numeric)-1
factor_mat <- as.matrix(factor_df)
Xmat_tr <- cbind(numeric_mat, factor_mat)
```

```
numeric_df <- X_te[, sapply(X_te, is.numeric)]
numeric_mat <- as.matrix(numeric_df)
factor_df <- sapply(X_te[, sapply(X_te, is.factor)], as.numeric)-1
factor_mat <- as.matrix(factor_df)
Xmat_te <- cbind(numeric_mat, factor_mat)
```

```
lasso_fit <- glmnet(Xmat_tr, as.numeric(Y_tr), family = "gaussian", alpha = 1)
```

```
cv.out <- cv.glmnet(Xmat_tr, as.numeric(Y_tr), family = "gaussian")
plot(cv.out)
```




```
bestlam <- cv.out$lambda.min
lasso_pred <- predict(lasso_fit , s = bestlam , newx = Xmat_te, type = "response")
yhat <- ifelse(lasso_pred >1, 1, 0)
table(yhat)
```

```
## yhat
##   0   1
## 434  11
```

```
conf.mat <- table(yhat, Y_te)
accuracy <- sum(diag(conf.mat))/sum(conf.mat)
accuracy
```

```
## [1] 0.7932584
```

```
print(bestlam)# best lambda roughly 0.03, 80 accuracy
```

```
## [1] 0.03198087
```

```
all_coef <- coef(lasso_fit, s = bestlam)
nonzero_coef <- all_coef[all_coef[, 1] != 0, ]
cat("Non-zero coefficients:\n")
```

Extract non-zero variables

```
## Non-zero coefficients:
```

```
nonzero_coef
```

```
##                        (Intercept)                 Num.of.pregnancies
##                        0.308724415                       0.090557018
##                    Smokes..years.                Smokes..packs.year.
##                        0.250750023                      -0.177074166
```

```
##    Hormonal.Contraceptives..years.                            IUD
##                        0.404512587                    0.111749718
##               STDs.condylomatosis      STDs.vaginal.condylomatosis
##                        0.294239141                   -0.419440231
## STDs.vulvo.perineal.condylomatosis                  STDs.syphilis
##                        0.056898940                   -0.049475011
##               STDs.genital.herpes                       STDs.HIV
##                        0.005797908                    0.680771219
##                  STDs.Hepatitis.B
##                       -0.378450200
```

4 Model training

linear regression

```
# Function to fit models for a given dataset
linear_1_2 <- lm(Outcome ~ Num.of.pregnancies + Smokes..years. + Smokes..packs.year. +Hormonal.Contracep
STDs.syphilis + STDs.genital.herpes + STDs.HIV + STDs.Hepatitis.B, data = cervical_tr_1_2)
linear_1_3 <- lm(Outcome ~ Num.of.pregnancies + Smokes..years. + Smokes..packs.year. +Hormonal.Contracep
STDs.syphilis + STDs.genital.herpes + STDs.HIV + STDs.Hepatitis.B, data = cervical_tr_1_3)
linear_1_4 <- lm(Outcome ~ Num.of.pregnancies + Smokes..years. + Smokes..packs.year. +Hormonal.Contracep
STDs.syphilis + STDs.genital.herpes + STDs.HIV + STDs.Hepatitis.B, data = cervical_tr_1_4)
linear_1_5 <- lm(Outcome ~ Num.of.pregnancies + Smokes..years. + Smokes..packs.year. +Hormonal.Contracep
STDs.syphilis + STDs.genital.herpes + STDs.HIV + STDs.Hepatitis.B, data = cervical_tr_1_5)
```

predict

```
linear_pred_1_2 <- predict(linear_1_2, cervical_te_1_2)
```

```
## Warning in predict.lm(linear_1_2, cervical_te_1_2): prediction from a
## rank-deficient fit may be misleading
```

```
linear_pred_1_3 <- predict(linear_1_3, cervical_te_1_3)
```

```
## Warning in predict.lm(linear_1_3, cervical_te_1_3): prediction from a
## rank-deficient fit may be misleading
```

```
linear_pred_1_4 <- predict(linear_1_4, cervical_te_1_4)
```

```
## Warning in predict.lm(linear_1_4, cervical_te_1_4): prediction from a
## rank-deficient fit may be misleading
```

```
linear_pred_1_5 <- predict(linear_1_5, cervical_te_1_5)
```

```
## Warning in predict.lm(linear_1_5, cervical_te_1_5): prediction from a
## rank-deficient fit may be misleading
```

MSE

```
# Calculate MSE for each dataset
mse_1_2 <- mean((cervical_te_1_2$Outcome - linear_pred_1_2)^2)
mse_1_3 <- mean((cervical_te_1_3$Outcome - linear_pred_1_3)^2)
mse_1_4 <- mean((cervical_te_1_4$Outcome - linear_pred_1_4)^2)
mse_1_5 <- mean((cervical_te_1_5$Outcome - linear_pred_1_5)^2)
```

```r
# Print MSE for each dataset
cat("MSE for 1:2 dataset:\n")
```

```
## MSE for 1:2 dataset:
```

```r
print(mse_1_2)
```

```
## [1] 1.117185
```

```r
cat("MSE for 1:3 dataset:\n")
```

```
## MSE for 1:3 dataset:
```

```r
print(mse_1_3)
```

```
## [1] 0.9070979
```

```r
cat("MSE for 1:4 dataset:\n")
```

```
## MSE for 1:4 dataset:
```

```r
print(mse_1_4)
```

```
## [1] 0.787125
```

```r
cat("MSE for 1:5 dataset:\n")
```

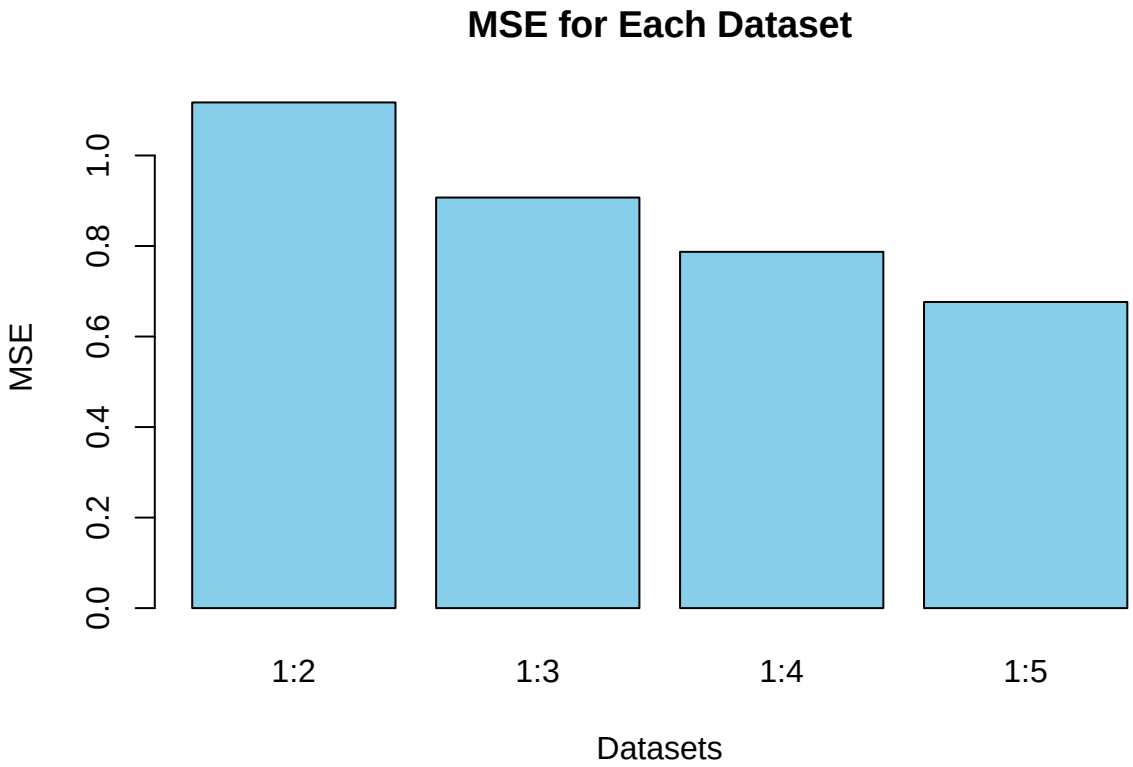
```
## MSE for 1:5 dataset:
```

```r
print(mse_1_5)
```

```
## [1] 0.6763406
```

plot MSE for each

```r
# Create a vector with MSE values
mse_values <- c(mse_1_2, mse_1_3, mse_1_4, mse_1_5)

# Create a vector with dataset names
datasets <- c("1:2", "1:3", "1:4", "1:5")

# Create a barplot for MSE values
barplot(mse_values, names.arg = datasets, xlab = "Datasets", ylab = "MSE", main = "MSE for Each Dataset
```



Do the same for random forest model

```
# Fit random forest models
library(randomForest)
```

```
## randomForest 4.7-1.1
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
##
## Attaching package: 'randomForest'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
## The following object is masked from 'package:ggplot2':
##
##     margin
```

```
rf_1_2 <- randomForest(Outcome ~ Hormonal.Contraceptives..years. + IUD..years. + STDs..Number.of.diagnos
```

```
## Warning in randomForest.default(m, y, ...): The response has five or fewer
## unique values.  Are you sure you want to do regression?
```

```
rf_1_3 <- randomForest(Outcome ~ Hormonal.Contraceptives..years. + IUD..years. + STDs..Number.of.diagnos
```

```
## Warning in randomForest.default(m, y, ...): The response has five or fewer
## unique values.  Are you sure you want to do regression?
```

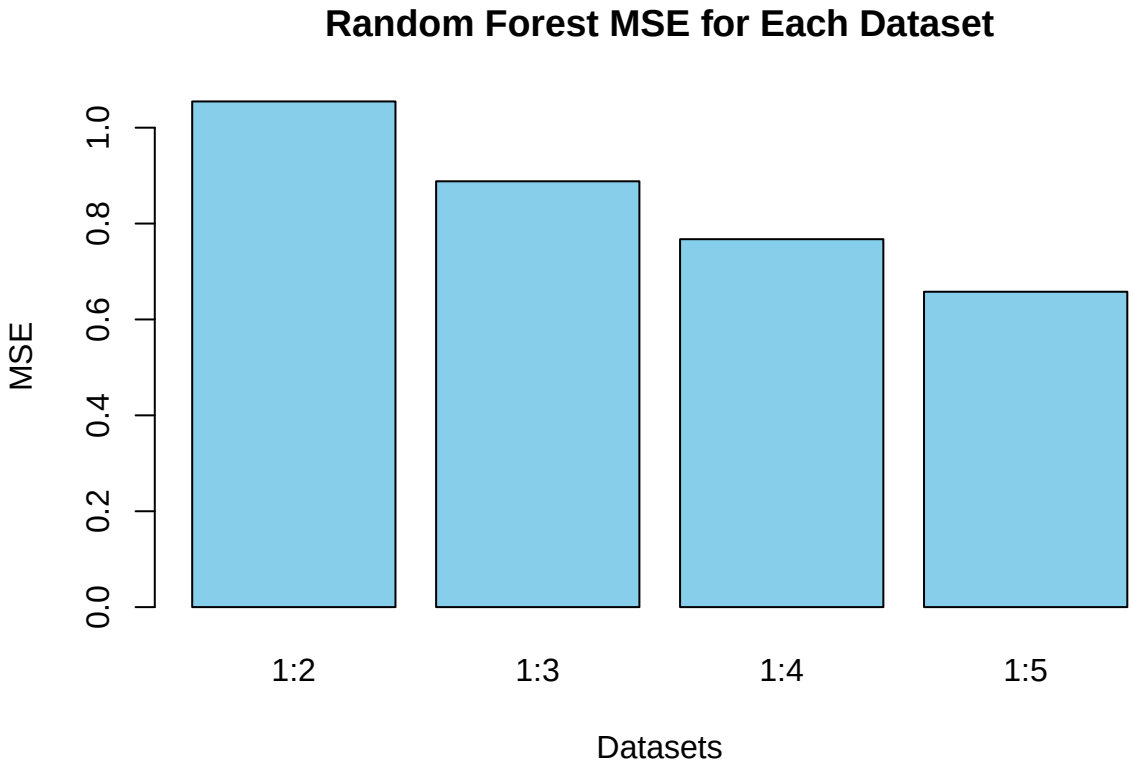
```
rf_1_4 <- randomForest(Outcome ~ Hormonal.Contraceptives..years. + IUD..years. + STDs..Number.of.diagnos
```

```
## Warning in randomForest.default(m, y, ...): The response has five or fewer
## unique values.  Are you sure you want to do regression?
```

```
rf_1_5 <- randomForest(Outcome ~ Hormonal.Contraceptives..years. + IUD..years. + STDs..Number.of.diagnos
```

```
## Warning in randomForest.default(m, y, ...): The response has five or fewer
## unique values.  Are you sure you want to do regression?
```

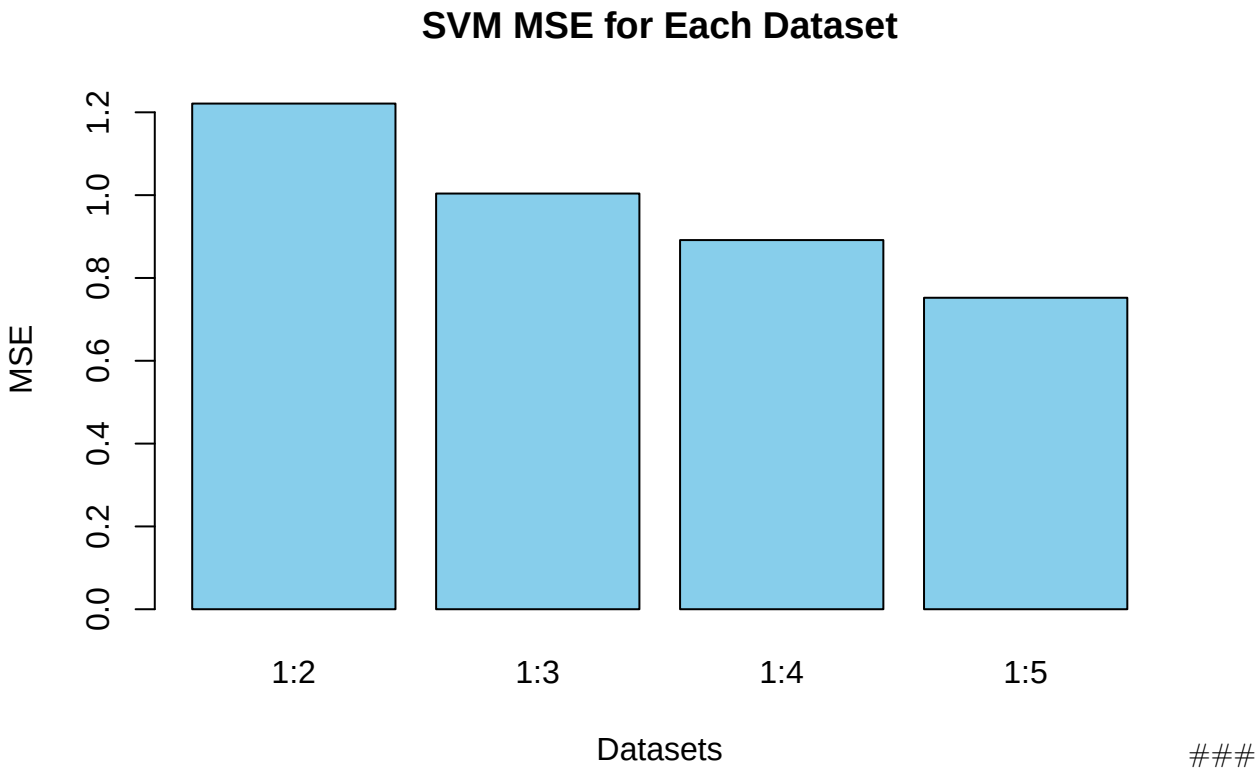
```
# Make predictions
rf_pred_1_2 <- predict(rf_1_2, cervical_te_1_2)
rf_pred_1_3 <- predict(rf_1_3, cervical_te_1_3)
rf_pred_1_4 <- predict(rf_1_4, cervical_te_1_4)
rf_pred_1_5 <- predict(rf_1_5, cervical_te_1_5)

# Calculate MSE
mse_rf_1_2 <- mean((cervical_te_1_2$Outcome - rf_pred_1_2)^2)
mse_rf_1_3 <- mean((cervical_te_1_3$Outcome - rf_pred_1_3)^2)
mse_rf_1_4 <- mean((cervical_te_1_4$Outcome - rf_pred_1_4)^2)
mse_rf_1_5 <- mean((cervical_te_1_5$Outcome - rf_pred_1_5)^2)

# Create a vector with MSE values
mse_rf_values <- c(mse_rf_1_2, mse_rf_1_3, mse_rf_1_4, mse_rf_1_5)

# Create a bar plot for MSE values
barplot(mse_rf_values, names.arg = datasets, xlab = "Datasets", ylab = "MSE", main = "Random Forest MSE
```



Do the same for the SVM

```
# Load the necessary library
library(e1071)

# Fit SVM models
svm_1_2 <- svm(Outcome ~ Hormonal.Contraceptives..years. + IUD..years. + STDs..Number.of.diagnosis + IU
```

```
svm_1_3 <- svm(Outcome ~ Hormonal.Contraceptives..years. + IUD..years. + STDs..Number.of.diagnosis + IU
svm_1_4 <- svm(Outcome ~ Hormonal.Contraceptives..years. + IUD..years. + STDs..Number.of.diagnosis + IU
svm_1_5 <- svm(Outcome ~ Hormonal.Contraceptives..years. + IUD..years. + STDs..Number.of.diagnosis + IU

# Make predictions
svm_pred_1_2 <- predict(svm_1_2, cervical_te_1_2)
svm_pred_1_3 <- predict(svm_1_3, cervical_te_1_3)
svm_pred_1_4 <- predict(svm_1_4, cervical_te_1_4)
svm_pred_1_5 <- predict(svm_1_5, cervical_te_1_5)

# Calculate MSE
mse_svm_1_2 <- mean((cervical_te_1_2$Outcome - svm_pred_1_2)^2)
mse_svm_1_3 <- mean((cervical_te_1_3$Outcome - svm_pred_1_3)^2)
mse_svm_1_4 <- mean((cervical_te_1_4$Outcome - svm_pred_1_4)^2)
mse_svm_1_5 <- mean((cervical_te_1_5$Outcome - svm_pred_1_5)^2)

# Create a vector with MSE values
mse_svm_values <- c(mse_svm_1_2, mse_svm_1_3, mse_svm_1_4, mse_svm_1_5)

# Create a bar plot for MSE values
barplot(mse_svm_values, names.arg = datasets, xlab = "Datasets", ylab = "MSE", main = "SVM MSE for Each
```



###

gbm

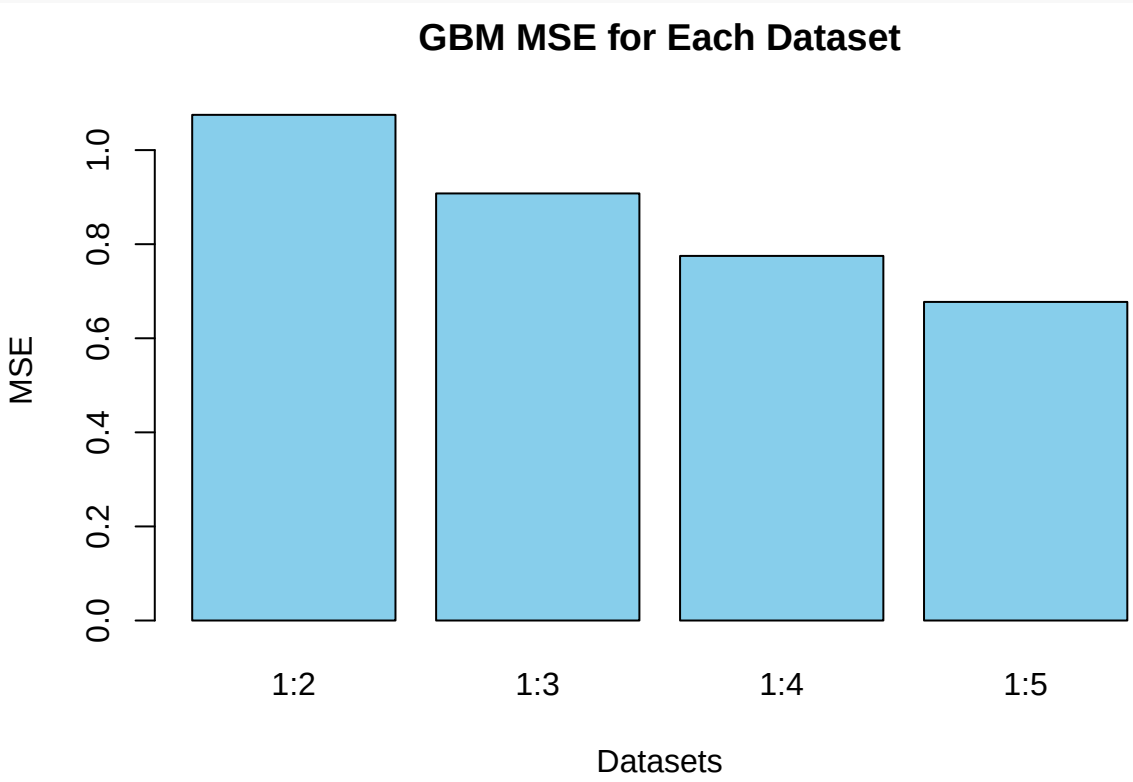
```
# Load the necessary libraries
library(gbm)
```

```
## Loaded gbm 2.1.8.1
```

```r
library(caret)
```

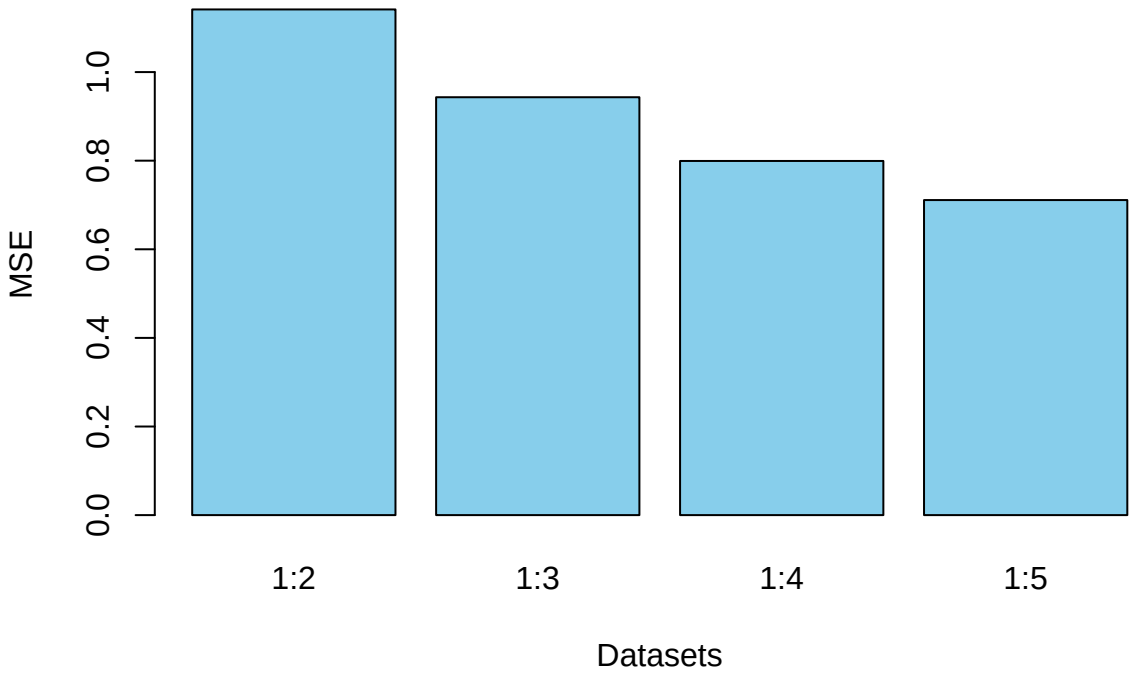
```
## Loading required package: lattice
```

```r
# Fit GBM models
gbm_1_2 <- gbm(Outcome ~ Hormonal.Contraceptives..years. + IUD..years. + STDs..Number.of.diagnosis + IU
gbm_1_3 <- gbm(Outcome ~ Hormonal.Contraceptives..years. + IUD..years. + STDs..Number.of.diagnosis + IU
gbm_1_4 <- gbm(Outcome ~ Hormonal.Contraceptives..years. + IUD..years. + STDs..Number.of.diagnosis + IU
gbm_1_5 <- gbm(Outcome ~ Hormonal.Contraceptives..years. + IUD..years. + STDs..Number.of.diagnosis + IU

# Make predictions
gbm_pred_1_2 <- predict(gbm_1_2, cervical_te_1_2, n.trees = 100)
gbm_pred_1_3 <- predict(gbm_1_3, cervical_te_1_3, n.trees = 100)
gbm_pred_1_4 <- predict(gbm_1_4, cervical_te_1_4, n.trees = 100)
gbm_pred_1_5 <- predict(gbm_1_5, cervical_te_1_5, n.trees = 100)

# Calculate MSE
mse_gbm_1_2 <- mean((cervical_te_1_2$Outcome - gbm_pred_1_2)^2)
mse_gbm_1_3 <- mean((cervical_te_1_3$Outcome - gbm_pred_1_3)^2)
mse_gbm_1_4 <- mean((cervical_te_1_4$Outcome - gbm_pred_1_4)^2)
mse_gbm_1_5 <- mean((cervical_te_1_5$Outcome - gbm_pred_1_5)^2)

# Create a vector with MSE values
mse_gbm_values <- c(mse_gbm_1_2, mse_gbm_1_3, mse_gbm_1_4, mse_gbm_1_5)

# Create a bar plot for MSE values
barplot(mse_gbm_values, names.arg = datasets, xlab = "Datasets", ylab = "MSE", main = "GBM MSE for Each
```



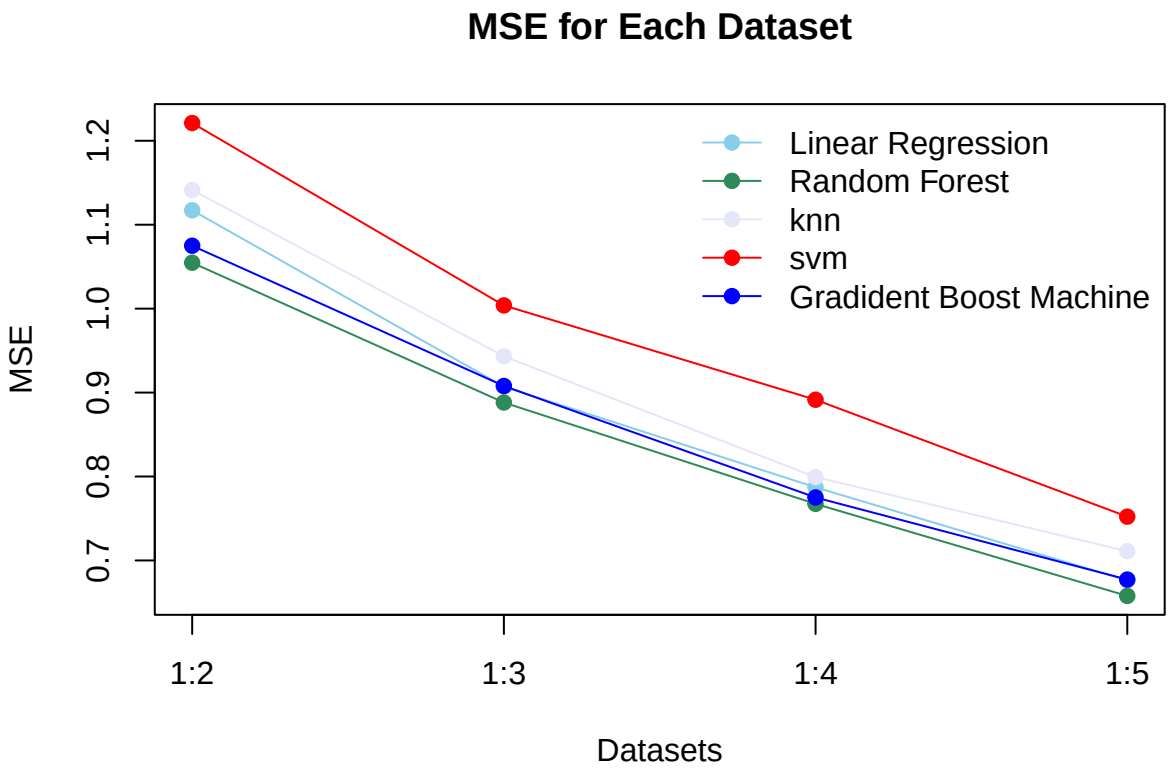

```r
# Load the necessary libraries
library(caret)
```

```r
# Prepare the data
train_1_2 <- cervical_tr_1_2[, c("Hormonal.Contraceptives..years.", "IUD..years.", "STDs..Number.of.dia
train_1_3 <- cervical_tr_1_3[, c("Hormonal.Contraceptives..years.", "IUD..years.", "STDs..Number.of.dia
train_1_4 <- cervical_tr_1_4[, c("Hormonal.Contraceptives..years.", "IUD..years.", "STDs..Number.of.dia
train_1_5 <- cervical_tr_1_5[, c("Hormonal.Contraceptives..years.", "IUD..years.", "STDs..Number.of.dia

# Specify the optimal value of k (number of neighbors)
k <- 5

# Fit knnreg models
knnreg_model_1_2 <- knnreg(train_1_2, cervical_te_1_2$Outcome, k)
knnreg_model_1_3 <- knnreg(train_1_3, cervical_te_1_3$Outcome, k)
knnreg_model_1_4 <- knnreg(train_1_4, cervical_te_1_4$Outcome, k)
knnreg_model_1_5 <- knnreg(train_1_5, cervical_te_1_5$Outcome, k)

# Make predictions
knnreg_pred_1_2 <- predict(knnreg_model_1_2, train_1_2)
knnreg_pred_1_3 <- predict(knnreg_model_1_3, train_1_3)
knnreg_pred_1_4 <- predict(knnreg_model_1_4, train_1_4)
knnreg_pred_1_5 <- predict(knnreg_model_1_5, train_1_5)

# Calculate MSE
mse_knnreg_1_2 <- mean((cervical_te_1_2$Outcome - knnreg_pred_1_2)^2)
mse_knnreg_1_3 <- mean((cervical_te_1_3$Outcome - knnreg_pred_1_3)^2)
mse_knnreg_1_4 <- mean((cervical_te_1_4$Outcome - knnreg_pred_1_4)^2)
mse_knnreg_1_5 <- mean((cervical_te_1_5$Outcome - knnreg_pred_1_5)^2)

# Create a vector with MSE values
mse_knnreg_values <- c(mse_knnreg_1_2, mse_knnreg_1_3, mse_knnreg_1_4, mse_knnreg_1_5)

# Create a bar plot for MSE values
barplot(mse_knnreg_values, names.arg = datasets, xlab = "Datasets", ylab = "MSE", main = "kNN Regressio
```

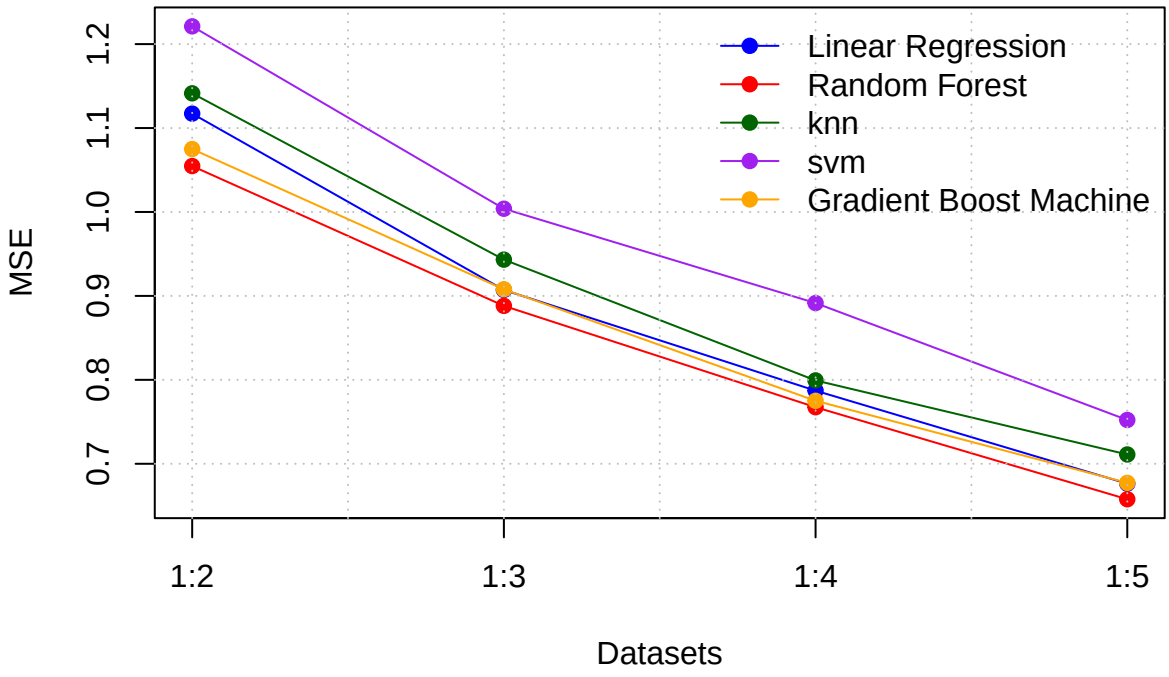



```r
# Create a line plot for MSE values of Linear Regression, Random Forest, and SVM
# Prepare the data
mse <- function(true, pred) {
  return(mean((true - pred)^2))
}
mse_values <- matrix(nrow = 5, ncol = 4)
mse_values[1,] <- c(mse(cervical_te_1_2$Outcome, linear_pred_1_2),
                    mse(cervical_te_1_3$Outcome, linear_pred_1_3),
                    mse(cervical_te_1_4$Outcome, linear_pred_1_4),
                    mse(cervical_te_1_5$Outcome, linear_pred_1_5))

mse_values[2,] <- c(mse(cervical_te_1_2$Outcome, rf_pred_1_2),
                    mse(cervical_te_1_3$Outcome, rf_pred_1_3),
                    mse(cervical_te_1_4$Outcome, rf_pred_1_4),
                    mse(cervical_te_1_5$Outcome, rf_pred_1_5))
mse_values[3,] <- c(mse(cervical_te_1_2$Outcome, knnreg_pred_1_2),
                    mse(cervical_te_1_3$Outcome, knnreg_pred_1_3),
                    mse(cervical_te_1_4$Outcome, knnreg_pred_1_4),
                    mse(cervical_te_1_5$Outcome, knnreg_pred_1_5))
mse_values[4,] <- c(mse(cervical_te_1_2$Outcome, svm_pred_1_2),
                    mse(cervical_te_1_3$Outcome, svm_pred_1_3),
                    mse(cervical_te_1_4$Outcome, svm_pred_1_4),
                    mse(cervical_te_1_5$Outcome, svm_pred_1_5))
mse_values[5,] <- c(mse(cervical_te_1_2$Outcome, gbm_pred_1_2),
                    mse(cervical_te_1_3$Outcome, gbm_pred_1_3),
                    mse(cervical_te_1_4$Outcome, gbm_pred_1_4),
                    mse(cervical_te_1_5$Outcome, gbm_pred_1_5))

datasets <- c("1:2", "1:3", "1:4", "1:5")
```

```r
# Create a line plot for MSE values of Linear Regression, Random Forest, and SVM
plot(mse_values[1,], type = "o", xaxt = "n", xlab = "Datasets", ylab = "MSE", main = "MSE for Each Datas
lines(mse_values[2,], type = "o", col = "#2E8B57", pch = 19)
lines(mse_values[3,], type = "o", col = "#E6E6FA", pch = 19)
lines(mse_values[4,], type = "o", col = "red", pch = 19)
lines(mse_values[5,], type = "o", col = "blue", pch = 19)

# Add axis labels
axis(1, at = 1:length(datasets), labels = datasets)

# Add legend
legend("topright", legend = c("Linear Regression", "Random Forest", "knn","svm","Gradident Boost Machine
```

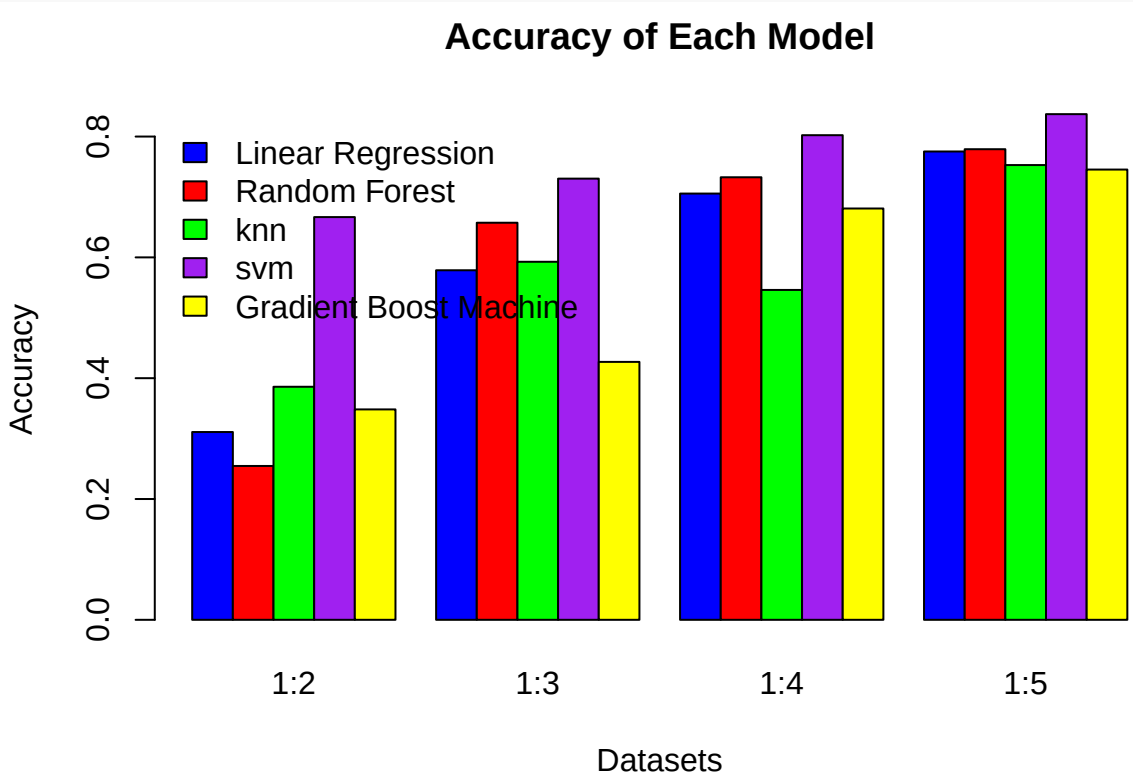



```r
library(grid)
# Create a line plot for MSE values of Linear Regression, Random Forest, and SVM
plot(mse_values[1,], type = "o", xaxt = "n", xlab = "Datasets", ylab = "MSE", main = "MSE for Each Datas
lines(mse_values[2,], type = "o", col = "red", pch = 19)
lines(mse_values[3,], type = "o", col = "darkgreen", pch = 19)
lines(mse_values[4,], type = "o", col = "purple", pch = 19)
lines(mse_values[5,], type = "o", col = "orange", pch = 19)

# Add axis labels
axis(1, at = 1:length(datasets), labels = datasets)

# Add gridlines
grid(nx = NULL, ny = NULL, col = "gray", lty = "dotted", lwd = par("lwd"), equilogs = TRUE)

# Add legend
legend("topright", legend = c("Linear Regression", "Random Forest", "knn", "svm", "Gradient Boost Machi
```




```r
# Calculate accuracy
accuracy <- function(true, pred) {
  return(mean(true == round(pred)))
}

acc_values <- matrix(nrow = 5, ncol = 4)
acc_values[1,] <- c(accuracy(cervical_te_1_2$Outcome, linear_pred_1_2),
                    accuracy(cervical_te_1_3$Outcome, linear_pred_1_3),
                    accuracy(cervical_te_1_4$Outcome, linear_pred_1_4),
                    accuracy(cervical_te_1_5$Outcome, linear_pred_1_5))

acc_values[2,] <- c(accuracy(cervical_te_1_2$Outcome, rf_pred_1_2),
                    accuracy(cervical_te_1_3$Outcome, rf_pred_1_3),
                    accuracy(cervical_te_1_4$Outcome, rf_pred_1_4),
                    accuracy(cervical_te_1_5$Outcome, rf_pred_1_5))

acc_values[3,] <- c(accuracy(cervical_te_1_2$Outcome, knnreg_pred_1_2),
                    accuracy(cervical_te_1_3$Outcome, knnreg_pred_1_3),
                    accuracy(cervical_te_1_4$Outcome, knnreg_pred_1_4),
                    accuracy(cervical_te_1_5$Outcome, knnreg_pred_1_5))

acc_values[4,] <- c(accuracy(cervical_te_1_2$Outcome, svm_pred_1_2),
                    accuracy(cervical_te_1_3$Outcome, svm_pred_1_3),
                    accuracy(cervical_te_1_4$Outcome, svm_pred_1_4),
                    accuracy(cervical_te_1_5$Outcome, svm_pred_1_5))

acc_values[5,] <- c(accuracy(cervical_te_1_2$Outcome, gbm_pred_1_2),
                    accuracy(cervical_te_1_3$Outcome, gbm_pred_1_3),
                    accuracy(cervical_te_1_4$Outcome, gbm_pred_1_4),
                    accuracy(cervical_te_1_5$Outcome, gbm_pred_1_5))
```

```r
# Create a grouped bar plot for accuracy values
barplot(acc_values, beside = TRUE, names.arg = datasets, xlab = "Datasets", ylab = "Accuracy", main = "
```



table

```r
# Create a data frame containing the accuracy values
accuracy_df <- data.frame(
  Model = rep(c("Linear Regression", "Random Forest", "knn","svm","Gradient Boost Machine"), each = 4),
  Dataset = rep(datasets, 5),
  Accuracy = c(acc_values)
)

# Round the accuracy values to 3 decimal places
accuracy_df$Accuracy <- round(accuracy_df$Accuracy, 3)

# Print the data frame
print(accuracy_df)
```

```
##                    Model Dataset Accuracy
## 1      Linear Regression     1:2    0.311
## 2      Linear Regression     1:3    0.255
## 3      Linear Regression     1:4    0.386
## 4      Linear Regression     1:5    0.667
## 5          Random Forest     1:2    0.348
## 6          Random Forest     1:3    0.579
## 7          Random Forest     1:4    0.657
## 8          Random Forest     1:5    0.593
## 9                    knn     1:2    0.730
## 10                   knn     1:3    0.427
```

```
## 11                    knn     1:4    0.706
## 12                    knn     1:5    0.733
## 13                    svm     1:2    0.546
## 14                    svm     1:3    0.802
## 15                    svm     1:4    0.681
## 16                    svm     1:5    0.775
## 17 Gradient Boost Machine     1:2    0.779
## 18 Gradient Boost Machine     1:3    0.753
## 19 Gradient Boost Machine     1:4    0.837
## 20 Gradient Boost Machine     1:5    0.745
```

Adjustments

Add interaction terms