R Notebook

Title: "IST687 – Text Mining HW"
Name: Sathish Kumar Rajendiran
Week: 8
Date: 05/27/2020

Exercise: Text Mining HW

Install necessary packages

```
install.packages( pkgs=c("gdata","RCurl","ggplot2","ggcorrplot","reshape2","ggeasy","viridis","viridisL
```

```
##
## The downloaded binary packages are in
##  /var/folders/_z/ltmjkt4156b37rsk7cgvj7l80000gn/T//Rtmp8vDfUj/downloaded_packages
```

```
library(gdata)
```

```
## gdata: read.xls support for 'XLS' (Excel 97-2004) files ENABLED.

##

## gdata: read.xls support for 'XLSX' (Excel 2007+) files ENABLED.

##
## Attaching package: 'gdata'

## The following object is masked from 'package:stats':
##
##     nobs

## The following object is masked from 'package:utils':
##
##     object.size

## The following object is masked from 'package:base':
##
##     startsWith
```

```
library(RCurl)
library(ggplot2)
library(ggcorrplot)
library(reshape2)
library(ggeasy)
library(viridis)
```

```
## Loading required package: viridisLite
```

```
library(viridisLite)
```

#printDataInfo function

```
# function printDataInfo
  printDataInfo <- function(myData)
    {
      strinfo <- str(myData)
      cat("str:",strinfo,"\n")

      colnamesinfo <- colnames(myData)
      cat("colnames:",colnamesinfo,"\n")

      diminfo <- dim(myData)
      cat("dim:",diminfo,"\n")

      nrowinfo <- nrow(myData)
      cat("nrow:",nrowinfo,"\n")

      nrowsinfo <- myData[1:5,]
      return(nrowsinfo)
     }
```

#1. Read in data from the following URL:

1. Read in data from the following URL:If you view this in a spreadsheet, you will find that four columns of a small dataset. The first column shows the number of fawn in a given spring (fawn are baby Antelope). The second column shows the population of adult antelope, the third shows the annual precipitation that year, and finally, the last column shows how bad the winter was during that year.
2. You have the option of saving the file save this file to your computer and read it into R, or reading the data directly from the web into a data frame.
3. You should inspect the data using the str() command to make sure that all of the cases have been read in (n=8 years of observations) and that there are four variables.

```
# filepath

    filepath <- "http://college.cengage.com/mathematics/brase/understandable_statistics/7e/students/data

     # function readFiles
     readFiles <- function(fpath)
        {
        dftemp <- read.xls(filepath)
        return(dftemp)
```

```
    }


    #Import the data into temporary datasets

    mlrdf <- data.frame(readFiles(filepath),stringsAsFactors = FALSE)
    mlrdf[1:3]
```

```
##    X1  X2   X3
## 1 2.9 9.2 13.2
## 2 2.4 8.7 11.5
## 3 2.0 7.2 10.8
## 4 2.3 8.5 12.3
## 5 3.2 9.6 12.6
## 6 1.9 6.8 10.6
## 7 3.4 9.7 14.1
## 8 2.1 7.9 11.2
```

```
    colnames(mlrdf) <- c("fawn","antelope","precipitation","wsi")
    #Review the dataframe
    printDataInfo(mlrdf)
```

```
## 'data.frame':    8 obs. of  4 variables:
##  $ fawn         : num  2.9 2.4 2 2.3 3.2 ...
##  $ antelope     : num  9.2 8.7 7.2 8.5 9.6 ...
##  $ precipitation: num  13.2 11.5 10.8 12.3 12.6 ...
##  $ wsi          : int  2 3 4 2 3 5 1 3
## str:
## colnames: fawn antelope precipitation wsi
## dim: 8 4
## nrow: 8
```

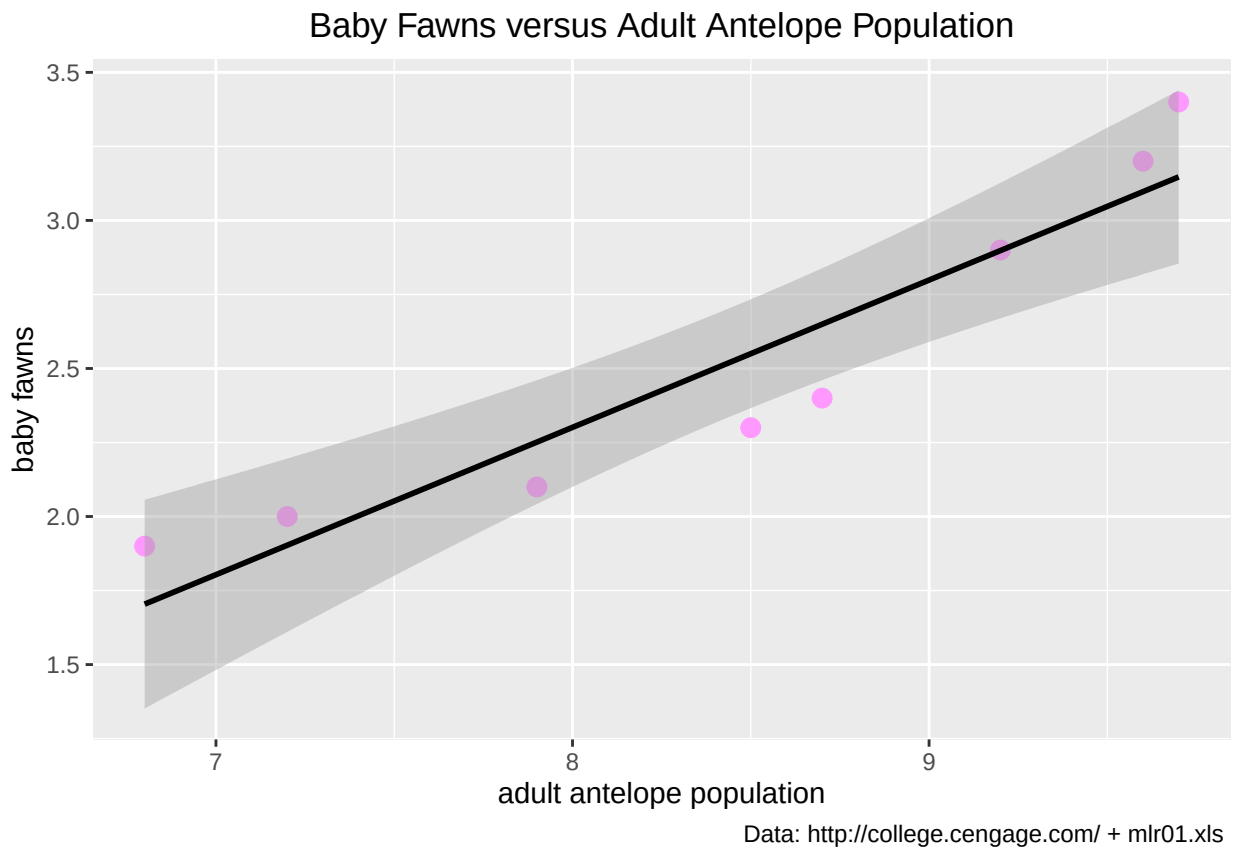
```
##   fawn antelope precipitation wsi
## 1  2.9      9.2          13.2   2
## 2  2.4      8.7          11.5   3
## 3  2.0      7.2          10.8   4
## 4  2.3      8.5          12.3   2
## 5  3.2      9.6          12.6   3
```

4. Create bivariate plots of number of baby fawns versus adult antelope population, the precipitation that year, and the severity of the winter. Your code should produce three separate plots. Make sure the Y-axis and X-axis are labeled. Keeping in mind that the number of fawns is the outcome (or dependent) variable, which axis should it go on in your plots?

```
#Baby Fawns versus Adult Antelope Population

  theme <-theme(plot.title = element_text(hjust = 0.5),axis.title = element_text())
  g1 <- ggplot(mlrdf,aes(x=antelope,y=fawn))+geom_point(aes(),size = 3,color="#FF99FF") + geom_smooth(m
  g1 + labs(x = "adult antelope population", y = "baby fawns",
            title = "Baby Fawns versus Adult Antelope Population",
            caption = "Data: http://college.cengage.com/ + mlr01.xls ") + theme
```

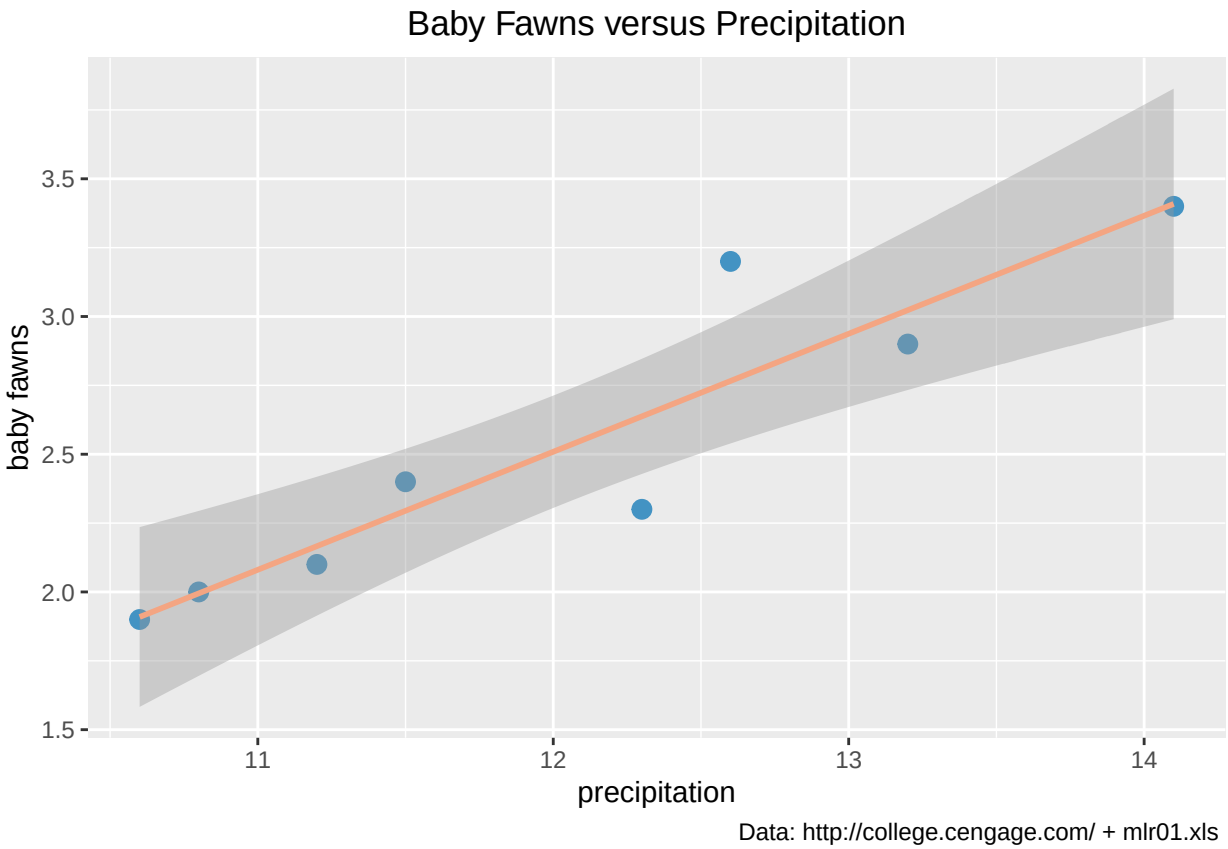
```
## `geom_smooth()` using formula 'y ~ x'
```




```
#Baby Fawns versus Precipitation

  g2 <- ggplot(mlrdf,aes(x=precipitation,y=fawn))+geom_point(aes(),size = 3,color="#4393C3") + geom_smo
  g2 + labs(x = "precipitation", y = "baby fawns",
            title = "Baby Fawns versus Precipitation",
            caption = "Data: http://college.cengage.com/ + mlr01.xls ") + theme
```

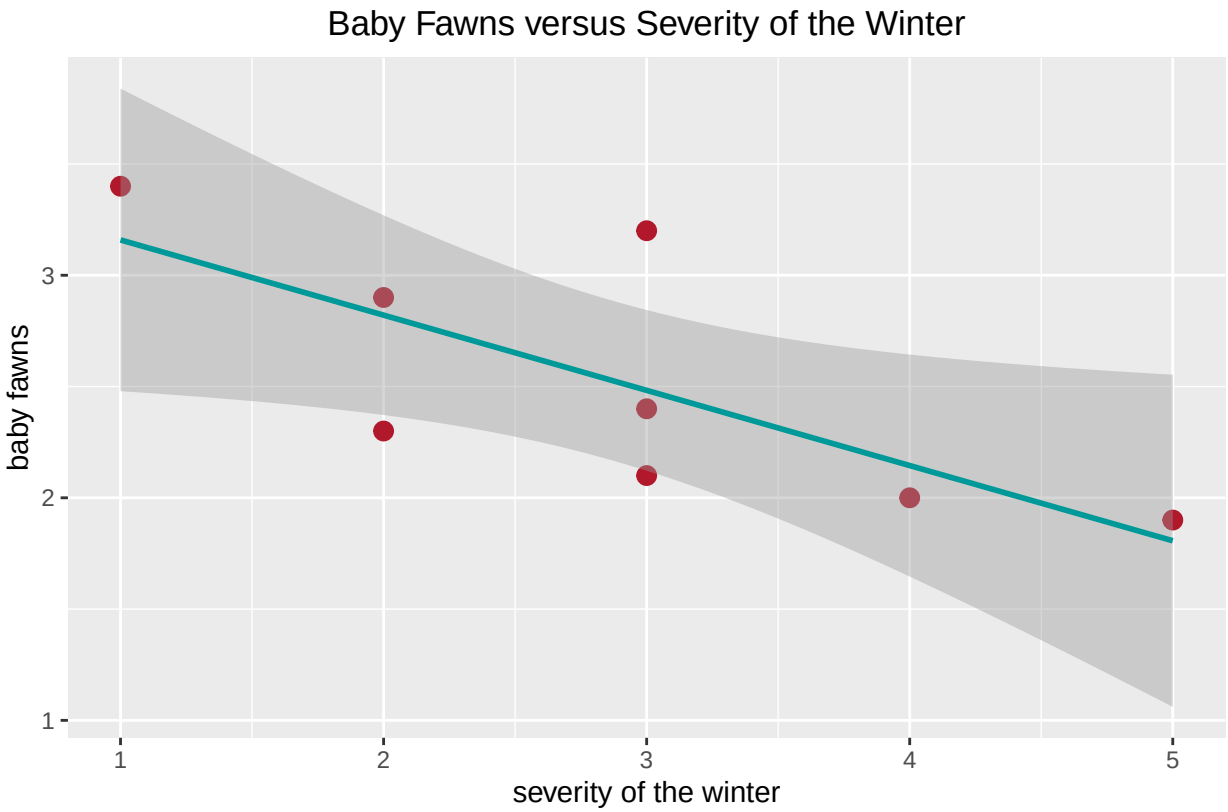
```
## `geom_smooth()` using formula 'y ~ x'
```




```
#Baby Fawns versus severity of the winter

 g3 <- ggplot(mlrdf,aes(x=wsi,y=fawn))+geom_point(aes(),size = 3,color="#B2182B") + geom_smooth(method
 g3 + labs(x = "severity of the winter", y = "baby fawns",
           title = "Baby Fawns versus Severity of the Winter",
           caption = "Data: http://college.cengage.com/ + mlr01.xls ") + theme
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Data: http://college.cengage.com/ + mlr01.xls

#5. Next, create three regression models of increasing complexity using lm()

```
    # predict the number of fawns from the severity of the winter

    mlr.lm1 <- lm(formula = fawn ~ wsi,data = mlrdf)
    summ <- summary(mlr.lm1)

    summary(mlr.lm1)
```

```
##
## Call:
## lm(formula = fawn ~ wsi, data = mlrdf)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.52069 -0.20431 -0.00172  0.13017  0.71724
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.4966     0.3904   8.957 0.000108 ***
## wsi          -0.3379     0.1258  -2.686 0.036263 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.415 on 6 degrees of freedom
## Multiple R-squared:  0.5459, Adjusted R-squared:  0.4702
## F-statistic: 7.213 on 1 and 6 DF,  p-value: 0.03626
```

```
    paste("p-values:")
```

```
## [1] "p-values:"
```

```
    summ$coefficients[,4]
```

```
## (Intercept)         wsi
## 0.000108158 0.036263036
```

```
    paste("adjusted r-squared:" ,summ$adj.r.squared)
```

```
## [1] "adjusted r-squared: 0.47020333641798"
```

```
    # str(summ)
```

```
    # predict the number of fawns from the severity of the winter + precipitation

    mlr.lm2 <- lm(formula = fawn ~ wsi+precipitation,data = mlrdf)
    summ <- summary(mlr.lm2)

    summary(mlr.lm2)
```

```
##
## Call:
## lm(formula = fawn ~ wsi + precipitation, data = mlrdf)
##
## Residuals:
##         1         2         3         4         5         6         7         8
## -0.165458  0.188313  0.006417 -0.193358  0.289080 -0.193312 -0.010695  0.079013
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -5.7791     2.2139  -2.610  0.04765 *
## wsi             0.2269     0.1490   1.522  0.18842
## precipitation   0.6357     0.1511   4.207  0.00843 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2133 on 5 degrees of freedom
## Multiple R-squared:    0.9,  Adjusted R-squared:   0.86
## F-statistic: 22.49 on 2 and 5 DF,  p-value: 0.003164
```

```
    paste("p-values:")
```

```
## [1] "p-values:"
```

```
    summ$coefficients[,4]
```

```
##   (Intercept)           wsi precipitation
##   0.047647946   0.188417206   0.008431877
```

```
    paste("adjusted r-squared:" ,summ$adj.r.squared)
```

```
## [1] "adjusted r-squared: 0.859962754071404"
```

```
    # str(summ)
```

```
   # predict the number of fawns from the severity of the winter + antelope

    mlr.lm3 <- lm(formula = fawn ~ wsi+antelope,data = mlrdf)
    summ <- summary(mlr.lm3)

    summary(mlr.lm3)
```

```
##
## Call:
## lm(formula = fawn ~ wsi + antelope, data = mlrdf)
##
## Residuals:
##       1       2       3       4       5       6       7       8
##  0.01231 -0.27531  0.10301 -0.19154  0.01535  0.15880  0.29992 -0.12256
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.46009    1.53443  -1.603   0.1698
## wsi          0.07058    0.12461   0.566   0.5956
## antelope     0.56594    0.14439   3.920   0.0112 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2252 on 5 degrees of freedom
## Multiple R-squared:  0.8885, Adjusted R-squared:  0.8439
## F-statistic: 19.92 on 2 and 5 DF,  p-value: 0.004152
```

```
    paste("p-values:")
```

```
## [1] "p-values:"
```

```
    summ$coefficients[,4]
```

```
## (Intercept)         wsi    antelope
##  0.16977988  0.59557987  0.01118699
```

```
    paste("adjusted r-squared:" ,summ$adj.r.squared)
```

```
## [1] "adjusted r-squared: 0.84389367311556"
```

```
    # str(summ)
```

```r
 # predict the number of fawns from the severity of the winter + precipitation + antelope

mlr.lm4 <- lm(formula = fawn ~ wsi+precipitation+antelope,data = mlrdf)
summ <- summary(mlr.lm4)

summary(mlr.lm4)
```

```
##
## Call:
## lm(formula = fawn ~ wsi + precipitation + antelope, data = mlrdf)
##
## Residuals:
##       1       2       3       4       5       6       7       8
## -0.11533 -0.02661  0.09882 -0.11723  0.02734 -0.04854  0.11715  0.06441
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -5.92201    1.25562  -4.716   0.0092 **
## wsi            0.26295    0.08514   3.089   0.0366 *
## precipitation  0.40150    0.10990   3.653   0.0217 *
## antelope       0.33822    0.09947   3.400   0.0273 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1209 on 4 degrees of freedom
## Multiple R-squared:  0.9743, Adjusted R-squared:  0.955
## F-statistic: 50.52 on 3 and 4 DF,  p-value: 0.001229
```

```r
paste("p-values:")
```

```
## [1] "p-values:"
```

```r
summ$coefficients[,4]
```

```
##   (Intercept)           wsi precipitation      antelope
##   0.009196072   0.036626174   0.021707219   0.027272444
```

```r
paste("adjusted r-squared:" ,summ$adj.r.squared)
```

```
## [1] "adjusted r-squared: 0.955004704934087"
```

```r
# Conclusion: predict the number of fawns from the severity of the winter + precipitation has better Ad
```