1. Explain why we could hypothesize that File could be used to predict Change and Hour. Use correlation analysis to validate this hypothesis.

   CODE: cor.test(file$File,file$Change)$estimate
   cor.test(file$File,file$Hour)$estimate

   RESULT: Result that I got were: 0.9592891

   Result that I got were: 0.6543844

   EXPLANATION: Since both the results are greater than 0.5, you can say that there are **strongly correlated**. Hence, you can use the attribute "File" to predict 'Change' and 'Hour'.

---

2. Analyze the linear regression models predicting Change and Hour using File as a single predictor. What model has better goodness of fit (i.e. higher R2)?

   CODE: foundchange=lm(Change~File,data=file)
   foundhour=lm(Hour~File,data=file)

   summary(foundchange)$r.squared
   summary(foundhour)$r.squared

   RESULT: 0.9202355
   0.4282189

   EXPLANATION: Since the value of foundchange is higher, the model built using Change will Have better goodness of fit, when compared.

---

3. Run cross-validation with each model. Report the mean absolute errors as the prediction accuracy. When running cross-validation, you should set the random seed with your A number. For example, if your Anumber is A1234, you call set.seed(1234)

   CODE:

```
cross.validation = function(form, dataset, iterations = 75, ratio = 0.10)
 {
      resultchange = rep(0,iterations)
      size = nrow(dataset) #total number of rows in the dataset
      test.size = ratio*size #number of rows in the testing set
      set.seed(02236773)
      for(i in 1:iterations)
      {

test.idx = sample.int(size,test.size)
test.data = dataset[test.idx, ]
```

```
    train.data = dataset[-test.idx, ]
    model =lm(form, data=train.data)# model is built using training data
    pred.result = predict(model,test.data)
    actualchange.result = test.data$Change
    resultchange[i] = mean(abs(actualchange.result - pred.result))  }
  resultchange
}

foundchange1=cross.validation(Change~File,file)
mean(foundchange1)
```

RESULT: 23.49387

CODE:

```
cross.validation1 = function(form, dataset, iterations = 75, ratio = 0.10)
{
  resulthour = rep(0,iterations)
  size = nrow(dataset) #total number of rows in the dataset
  test.size = ratio*size #number of rows in the testing set
  set.seed(02236773)
  for(i in 1:iterations)
  {

    test.idx = sample.int(size,test.size)
    test.data = dataset[test.idx, ]
    train.data = dataset[-test.idx, ]
    model =lm(form, data=train.data
    pred.result = predict(model,test.data)
    actualhour.result=test.data$Hour
    resulthour[i] = mean(abs(actualhour.result - pred.result))
  }
  resulthour
}

foundhour1=cross.validation1(Hour~File,file)
mean(foundhour1)
```

RESULT: 132.2213

EXPLANATION: When compared, the Change model has got less value which proves that it is the best.

---

4. 4. In addition to File, what other factors could we use for effort prediction: programming language, team size, type of software (e.g. web app, mobile app, desktop app, embedded systems, operating system...) Explain why such a factor has the predictive power?

   EXPLANATION: In my opinion, we could use team size and the type of software for effort prediction. As the size of the team grows, it will take less number of working hours to develop a software and since it can be used to predict the efforts. Additionally, previous project data can also be used to predict the future efforts required, in the same software environment.