2 Forest Fire Data

(a) Download the Forest Fire data

(b) Exploring the data:

i. How many rows are in this data set? How many columns? What do the rows and columns represent?

three are 517 rows and 13 columns

each row is an information of fires with burned area.

columns represents information about the fire. for example, X and Y are axis spatial coordinate within the park map, month and date of the fire, FFMC, DMC, DC indexes, temperature, relative humidity, wind speed, rain, and the burned area of the forest.

ii. Explain why the transformation $Y1 = \ln(1 + Y)$, where Y is the response variable is useful for this dataset. In the following, use Y1 as the new response variable.

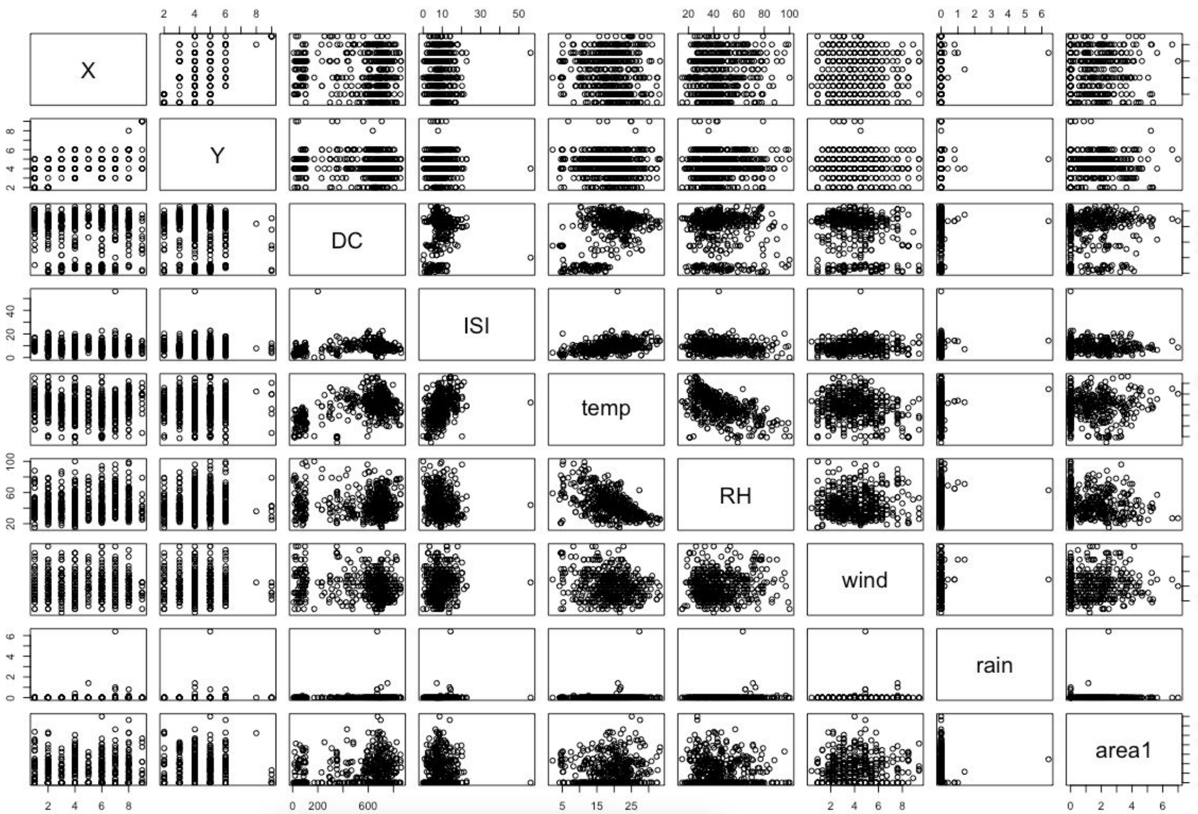
Y has many zero in the records. It has large standard deviation and is highly skewed,

It is hard to interpret any information from the distribution of Y

therefore we need to be transform Y in a format that

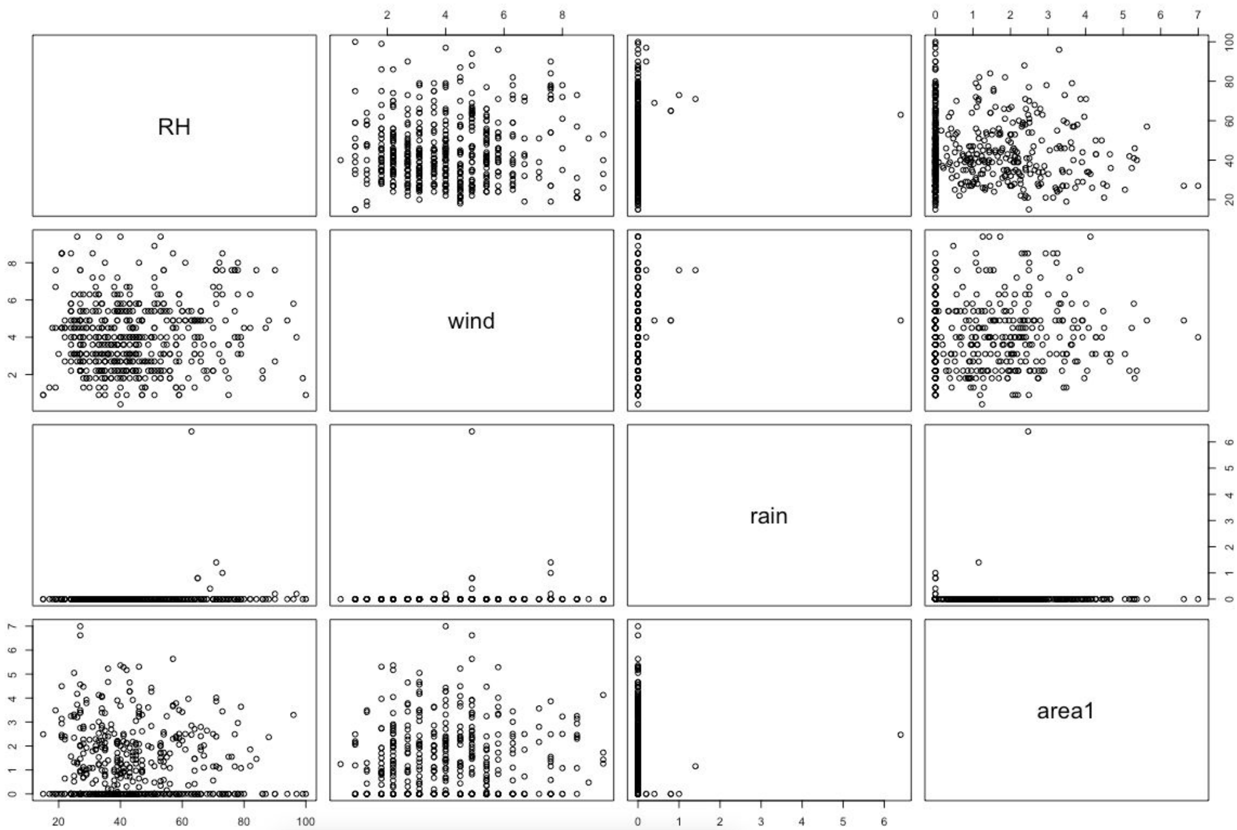
so needs to be transformed to log. and avoid the situation of log(0), we change to log(Y+1)

iii. Make pairwise scatterplots of the predictors (columns) in this data set with the dependent variable. Describe your findings.



According to the scatter plot, we can hardly find any linear relationship between independent variables and the dependent variable, area1. However, we can tell that there are some variables have some associations, such as temp and RH, ISI and RH.

iv. Make at least 16 pairwise scatterplots of predictors of your choice and describe your findings. You are welcome to make all possible scatter plots.



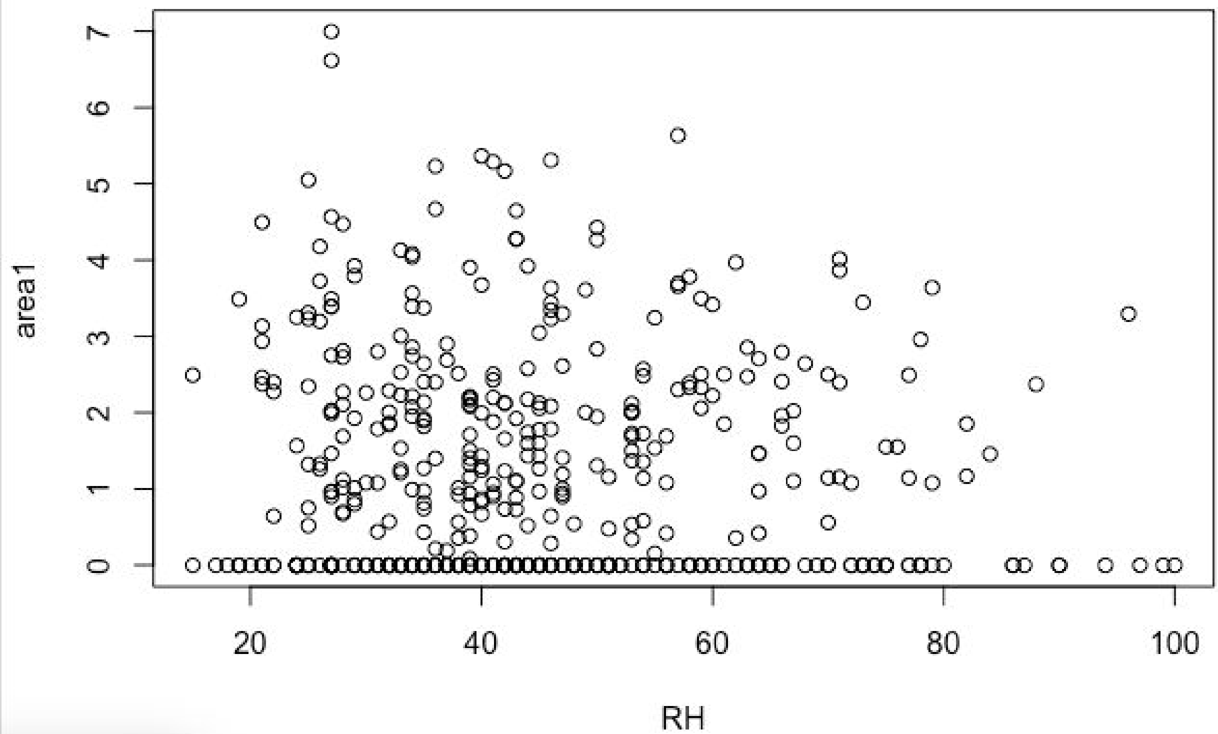
Initially, I thought there might be some linear relationship between area and rain, area and wind. However, the scatter plot disapproves my thoughts. There's no correlation between them. Therefore, we can tell the variables I pick have strong correlation with area.

v. What are the mean, the median, range, first and third quartiles, and interquartile ranges of each of the variables in the dataset? Summarize them in a table.

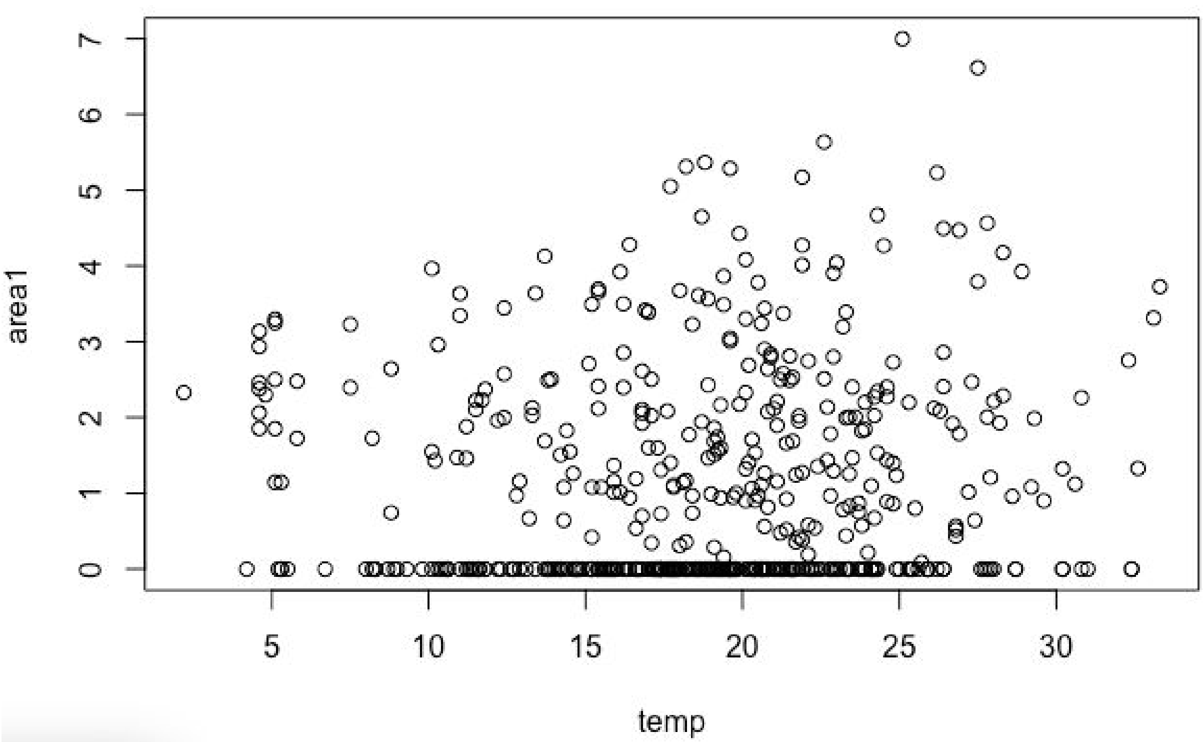
```
      X               Y             FFMC            DMC              DC             ISI              temp
Min.   :1.000   Min.   :2.0   Min.   :18.70   Min.   :  1.1   Min.   :  7.9   Min.   : 0.000   Min.   : 2.20
1st Qu.:3.000   1st Qu.:4.0   1st Qu.:90.20   1st Qu.: 68.6   1st Qu.:437.7   1st Qu.: 6.500   1st Qu.:15.50
Median :4.000   Median :4.0   Median :91.60   Median :108.3   Median :664.2   Median : 8.400   Median :19.30
Mean   :4.669   Mean   :4.3   Mean   :90.64   Mean   :110.9   Mean   :547.9   Mean   : 9.022   Mean   :18.89
3rd Qu.:7.000   3rd Qu.:5.0   3rd Qu.:92.90   3rd Qu.:142.4   3rd Qu.:713.9   3rd Qu.:10.800   3rd Qu.:22.80
Max.   :9.000   Max.   :9.0   Max.   :96.20   Max.   :291.3   Max.   :860.6   Max.   :56.100   Max.   :33.30
       RH              wind            rain              area1
Min.   : 15.00   Min.   :0.400   Min.   :0.00000   Min.   :0.0000
1st Qu.: 33.00   1st Qu.:2.700   1st Qu.:0.00000   1st Qu.:0.0000
Median : 42.00   Median :4.000   Median :0.00000   Median :0.4187
Mean   : 44.29   Mean   :4.018   Mean   :0.02166   Mean   :1.1110
3rd Qu.: 53.00   3rd Qu.:4.900   3rd Qu.:0.00000   3rd Qu.:2.0242
Max.   :100.00   Max.   :9.400   Max.   :6.40000   Max.   :6.9956
```

(c) Fit a simple linear regression.
Describe your results. In which of the models is there a statistically significant association between the predictor and the response? Create some plots to backup your assertions. Are there any outliers that you would like to remove from your data for each of these regression tasks?

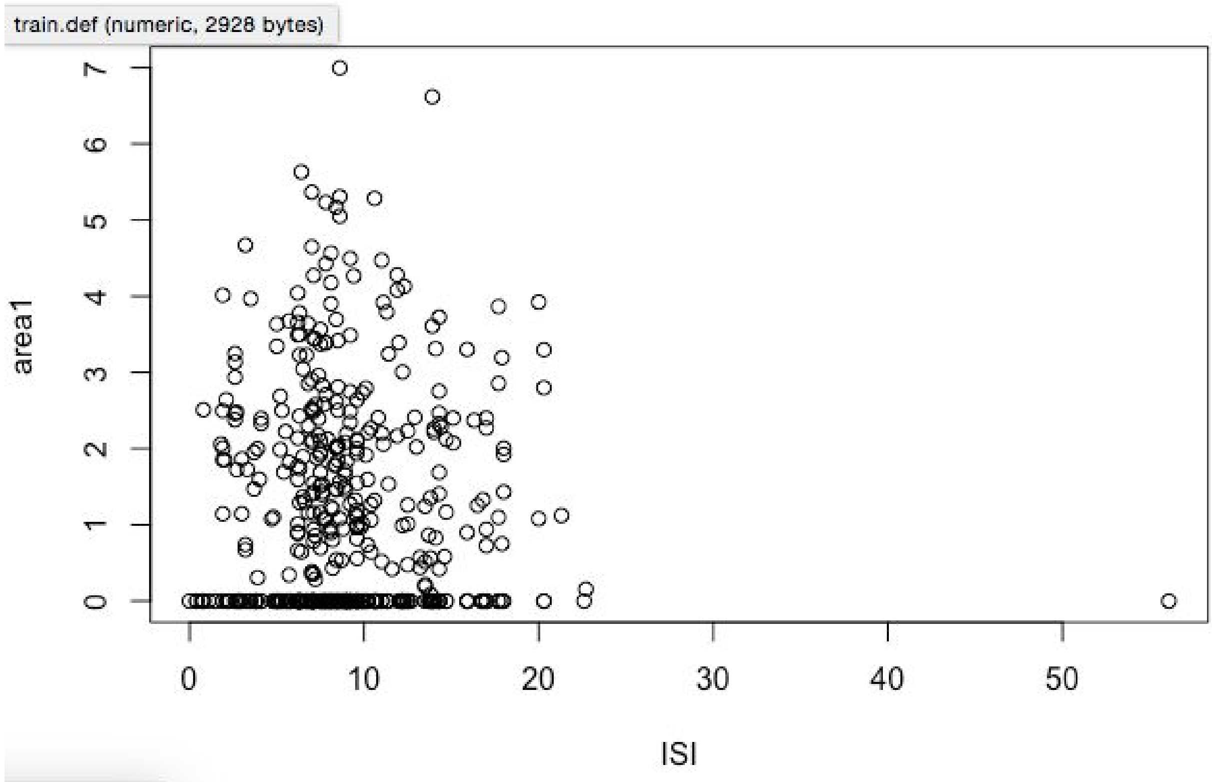
All variables have no statistically significant association between the predictor and the response. See some plots below:



area1
RH

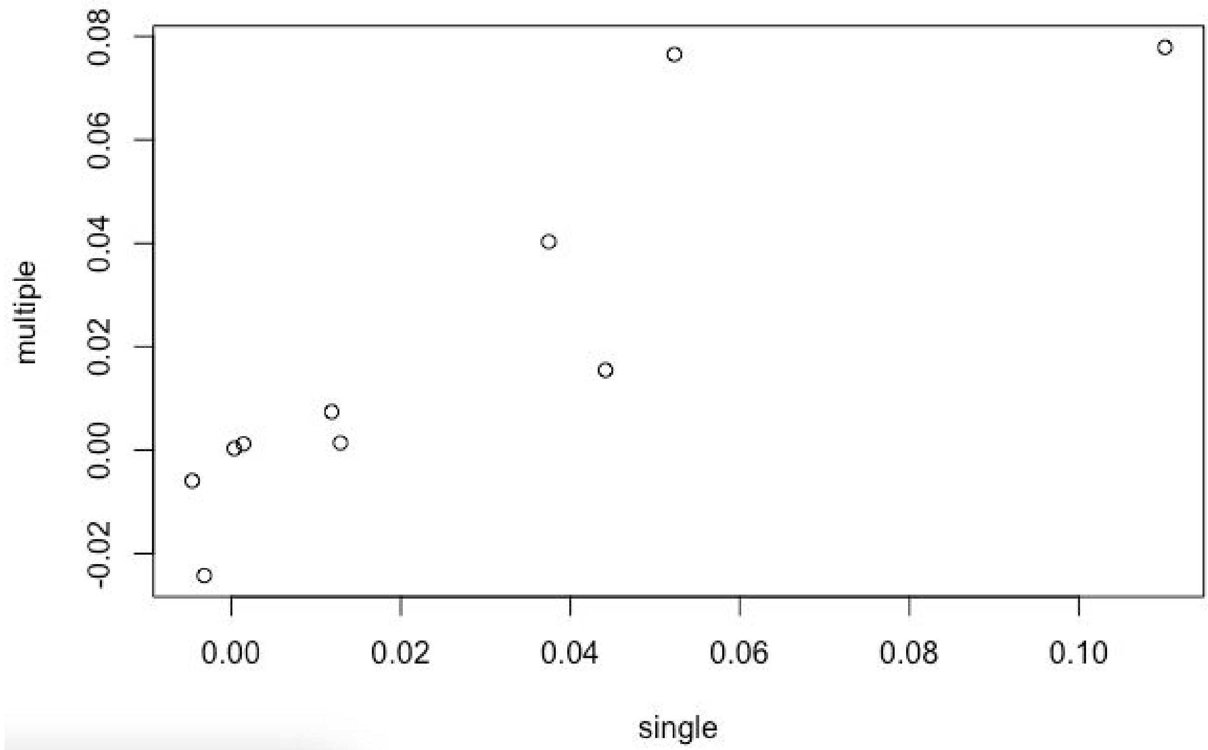


area1
temp



(d) Fit a multiple regression model to predict the response using all of the predictors. Describe your results. For which predictors can we reject the null hypothesis H0 :j =0? Wind is found to be statistically significant with the response. For other variables, we fail to reject the null hypothesis. R-squared is also the highest compared to single linear regression of each variable. more variance in predictor can be explained using this model.

(e) How do your results from 2c compare to your results from 2d? Create a plot displaying the univariate regression coefficients from 2c on the x-axis, and the multiple regression coefficients from 2d on the y-axis. That is, each predictor is displayed as a single point in the plot. Its coecient in a simple linear regression model is shown on the x-axis, and its coecient estimate in the multiple linear regression model is shown on the y-axis.



f) Is there evidence of nonlinear association between any of the predictors and the response? To answer this question, for each predictor X, fit a model of the form
Y1 =0 +1X+2X2 +3X3 +✏
Temp and wind have nonlinear association with the response.

(g) Is there evidence of association of interactions of predictors with the response? To answer this question, run a full linear regression model with all pairwise interaction terms and state whether any interaction terms are statistically significant.
Temp and DMC, temp and wind have statistically significant association of interactions with the response.

(h) Can you improve your model using possible interaction terms or nonlinear associations and between the predictors and response? Train the model on a randomly selected 70% subset of the data and test it on the remaining points and report your train and test results.

(i) KNN Regression: Note that for this problem, we have a mixture of categorical and quantitative predictors. There is not a unique way to define a distance metric in such a situation.

Describe your findings and heuristics. Can your metric be specific to this problem? Use a reasonable distance metric to answer the following questions:

i. Use the first 4 predictors to perform k-nearest neighbor regression for this dataset. Find the value of k that gives you the best fit. Plot the train and test errors in terms of 1/k.

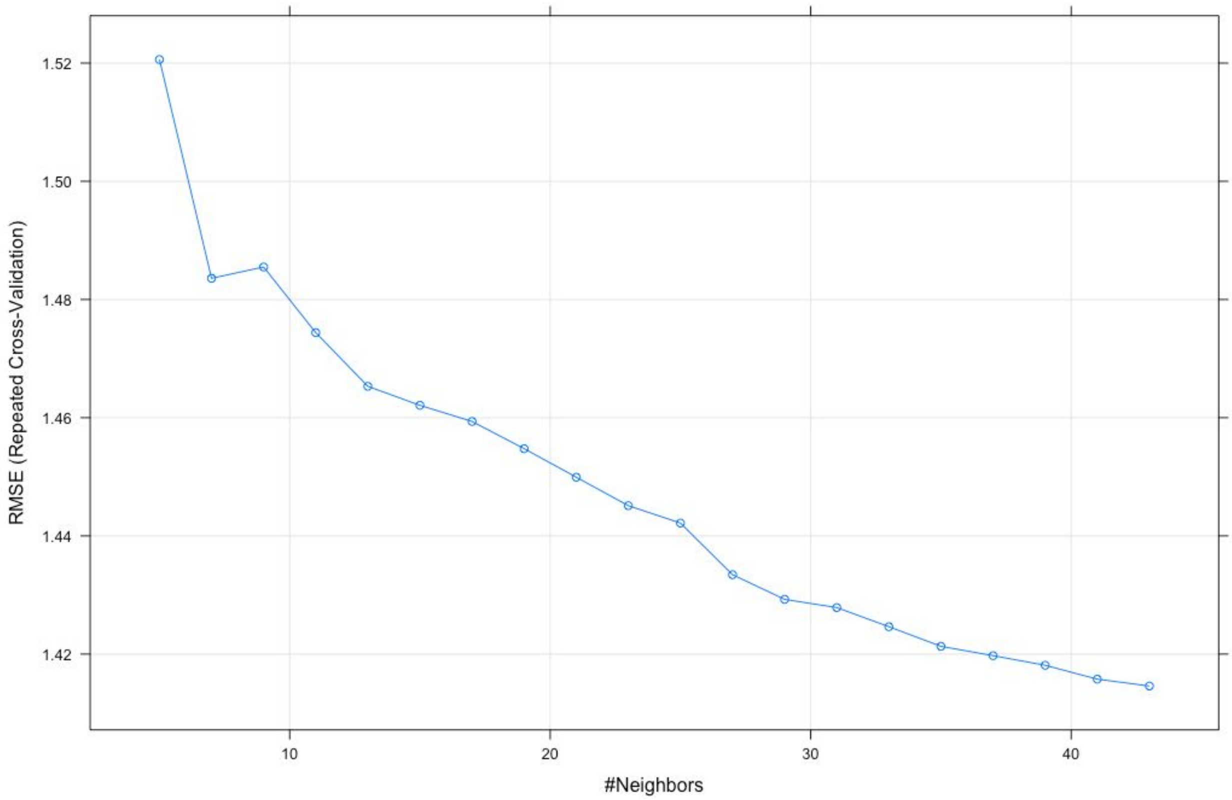
```
> knnFit
k-Nearest Neighbors

361 samples
  4 predictor

Pre-processing: centered (4), scaled (4)
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 325, 325, 325, 325, 325, 325, ...
Resampling results across tuning parameters:

  k   RMSE      Rsquared    RMSE SD    Rsquared SD
   5  1.520601  0.02723832  0.1594361  0.03242718
   7  1.483584  0.03335633  0.1523462  0.03566701
   9  1.485495  0.03314994  0.1490379  0.03381031
  11  1.474389  0.03887282  0.1475489  0.05055401
  13  1.465304  0.03652415  0.1528355  0.04747656
  15  1.462088  0.04243510  0.1567086  0.05726194
  17  1.459360  0.04472742  0.1527070  0.05647285
  19  1.454754  0.04102228  0.1503359  0.04683427
  21  1.449926  0.03579345  0.1486193  0.03974541
  23  1.445107  0.03965775  0.1488872  0.03892223
  25  1.442164  0.03352809  0.1451155  0.03514364
  27  1.433431  0.02517661  0.1454146  0.03262207
  29  1.429254  0.02129564  0.1411717  0.03410542
  31  1.427865  0.02293102  0.1423024  0.03310201
  33  1.424632  0.01999613  0.1412200  0.03093268
  35  1.421309  0.02055732  0.1413118  0.03397213
  37  1.419730  0.01867340  0.1426208  0.03117719
  39  1.418102  0.02050718  0.1422440  0.02827532
  41  1.415752  0.02096891  0.1406260  0.02608464
  43  1.414592  0.02043763  0.1389607  0.02498585

RMSE was used to select the optimal model using  the smallest value.
The final value used for the model was k = 43.
```



ii. Use the last 4 predictors to perform k-nearest neighbor regression for this dataset. Find the value of k that gives you the best fit. Plot the train and test errors in terms of 1/k.

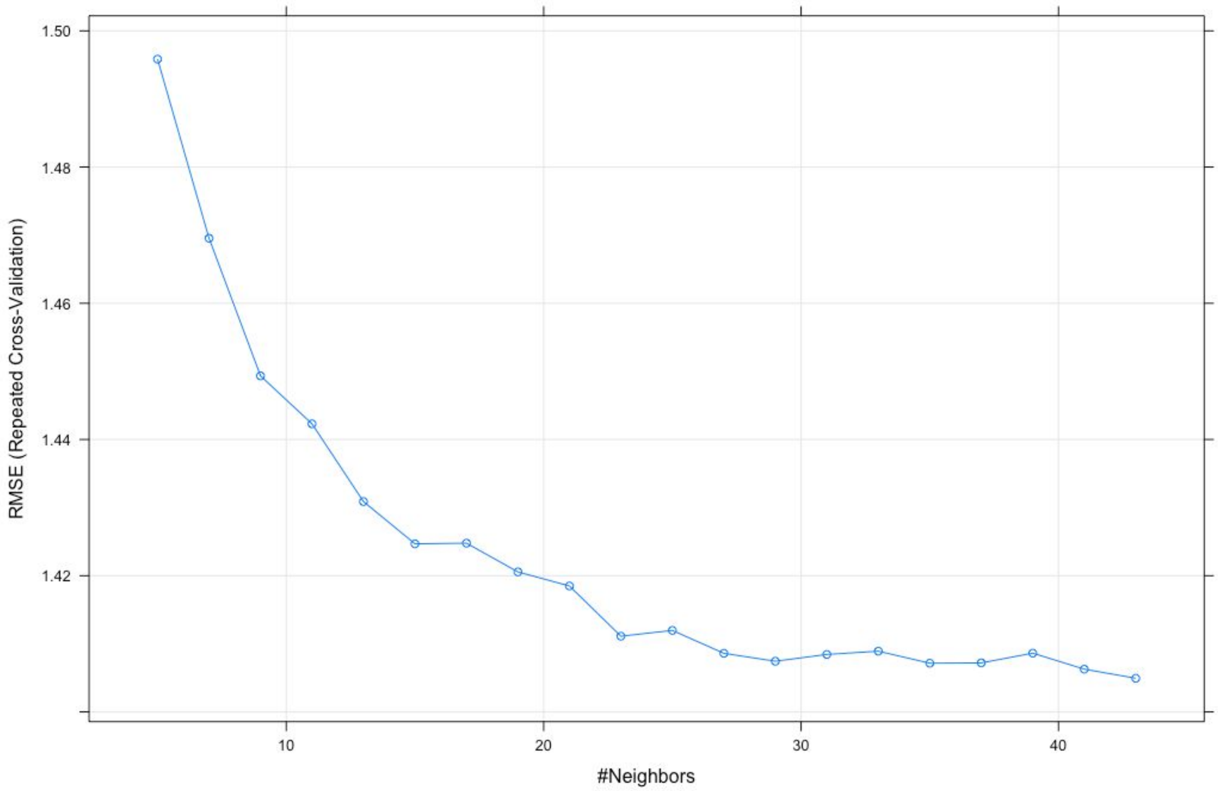
```
> knnFit2
set.seed(seed, kind = NULL, normal.kind = NULL)
k-Nearest Neighbors

361 samples
  4 predictor

Pre-processing: centered (4), scaled (4)
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 325, 324, 325, 325, 325, 325, ...
Resampling results across tuning parameters:

  k   RMSE      Rsquared    RMSE SD    Rsquared SD
   5  1.495847  0.02073121  0.1792292  0.02455284
   7  1.469547  0.01935895  0.1924418  0.03209807
   9  1.449341  0.01398230  0.1933601  0.01534646
  11  1.442297  0.01548130  0.1978107  0.01863266
  13  1.430882  0.01673099  0.1970731  0.02532658
  15  1.424682  0.01789289  0.1974178  0.02550042
  17  1.424777  0.01946519  0.1907213  0.03109871
  19  1.420556  0.02144034  0.1857199  0.03414100
  21  1.418492  0.01998484  0.1828784  0.03359545
  23  1.411117  0.01757381  0.1844890  0.02554407
  25  1.411966  0.01681264  0.1827669  0.02294452
  27  1.408604  0.01915406  0.1841034  0.02638275
  29  1.407440  0.01536410  0.1873329  0.02089620
  31  1.408437  0.01425129  0.1859655  0.01458351
  33  1.408915  0.01504708  0.1877619  0.01615123
  35  1.407151  0.01232108  0.1874852  0.01664176
  37  1.407197  0.01448281  0.1878104  0.01803976
  39  1.408625  0.01454557  0.1850155  0.01619025
  41  1.406278  0.01545536  0.1868074  0.01826695
  43  1.404939  0.01714027  0.1882537  0.02258634

RMSE was used to select the optimal model using  the smallest value.
The final value used for the model was k = 43.
```



iii. Use predictors 1,2, 9, 10, 11 to perform k-nearest neighbor regression for this dataset. Find the value of k that gives you the best fit. Plot the train and test errors in terms of 1/k.

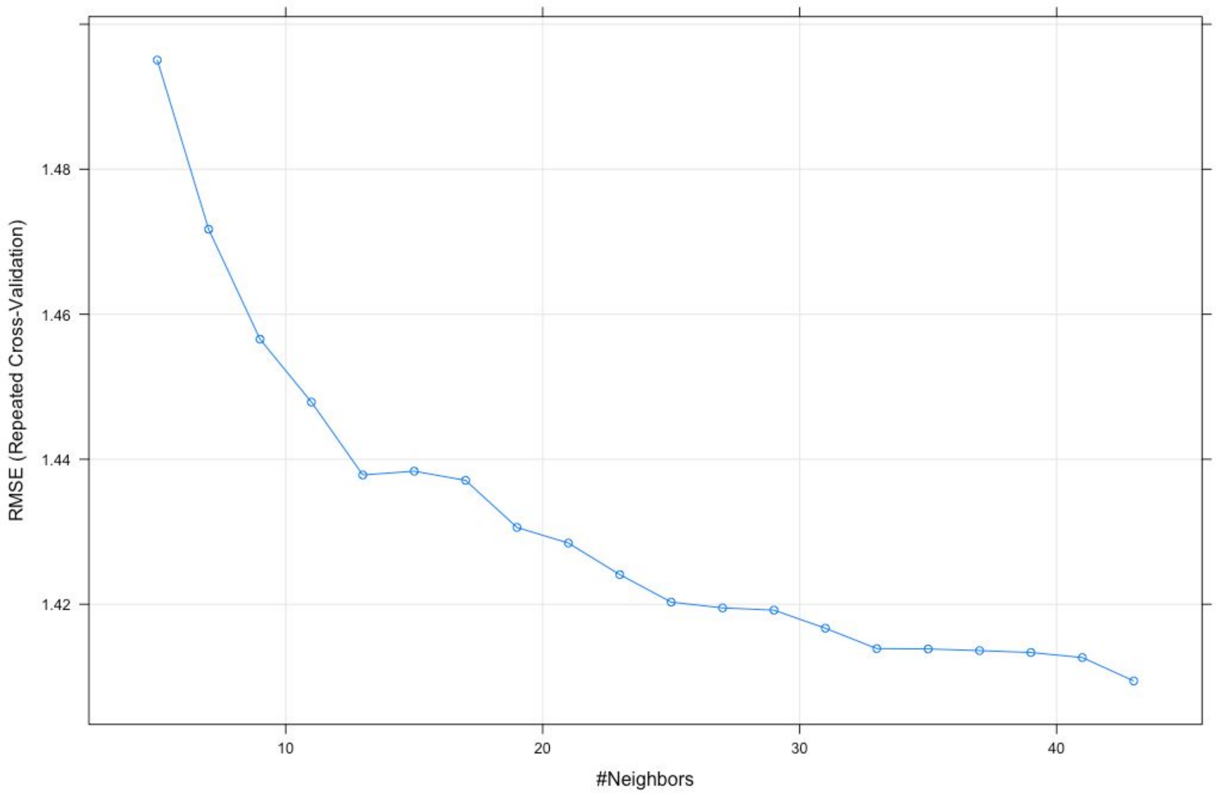
```
> knnFit3
k-Nearest Neighbors

361 samples
  4 predictor

Pre-processing: centered (4), scaled (4)
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 325, 325, 325, 325, 325, 325, ...
Resampling results across tuning parameters:

  k   RMSE      Rsquared    RMSE SD    Rsquared SD
   5  1.495057  0.04116353  0.1737052  0.05159216
   7  1.471718  0.03082662  0.1663313  0.04046160
   9  1.456552  0.03220335  0.1704165  0.04653463
  11  1.447883  0.03542402  0.1636147  0.05313705
  13  1.437831  0.03666064  0.1707206  0.05588575
  15  1.438351  0.03552657  0.1724325  0.05119310
  17  1.437085  0.03443697  0.1705062  0.05284073
  19  1.430596  0.03132480  0.1733009  0.04404385
  21  1.428438  0.03054187  0.1726119  0.03919322
  23  1.424089  0.02348228  0.1752891  0.03475019
  25  1.420297  0.02388572  0.1762360  0.03787216
  27  1.419497  0.02321015  0.1750823  0.03304729
  29  1.419189  0.01902266  0.1788007  0.03278247
  31  1.416694  0.02007139  0.1798918  0.02882327
  33  1.413883  0.02002697  0.1790392  0.02690677
  35  1.413833  0.02251003  0.1781193  0.02679848
  37  1.413605  0.02302066  0.1786510  0.02775247
  39  1.413344  0.02284562  0.1804659  0.02776154
  41  1.412648  0.02394096  0.1808099  0.02709781
  43  1.409423  0.02086493  0.1784842  0.02281711

RMSE was used to select the optimal model using  the smallest value.
The final value used for the model was k = 43.
```


RMSE (Repeated Cross-Validation)
1.48
1.46
1.44
1.42
10
20
30
40
#Neighbors