renuka

Renuka_Practicum2_1and2question

2023-11-15

install.packages("magrittr") install.packages("tinytex")

#1.(0 pts / 10 min) Download the data set Census Income Data for Adults along with its explanation. There are two data sets (adult.data and adult.test) along with a file describing the features (adult.names). Note that the data file does not contain header names; you may wish to add those to the data frame during reading. The description of each column can be found in the file adult.names. Explore the combined data set as you see fit and that allows you to get a sense of the data and get comfortable with it, but do not include any of the code for exploration in your notebook; perhaps create a new notebook for exploration.

#2.Load both data sets (adult.data and adult.test) and combine the two data sets into a single data frame that has headers. Display rows 11, 112, 199, and 203 and the first 4 columns of the combined data frame using the function head(). That way we can see if you loaded the data correctly and the columns have headers.

```
## # A tibble: 4 x 4
##     age workclass fnlwgt education
##   <dbl> <chr>      <dbl> <chr>
## 1    37 Private   280464 Some-college
## 2    38 Private    65324 Prof-school
## 3    35 Private   138992 Masters
## 4    51 Private   259323 Bachelors
```

#3. (4 pts / 20 min) Split the combined data set 75/25% so you retain 25% for validation using random sampling without replacement. Use a fixed seed of 33452, so you produce the same results each time you run the code. Going forward you will use the 75% data set for training and the 25% data set for validation and determining accuracy metrics.

```
## Train set size: 36632
```

```
## Validate set size: 12210
```

#4. (8 pts / 60 min) Using the Naive Bayes Classification algorithm from the KlaR package, build a binary classifier that predicts whether an individual earns more than or less than US$50k (the last column). Only use the features in column 1, 2, 5, 9, 10, 13, 14. Ignore any other features in your model. You need to transform continuous variables into categorical variables by binning (use equal size bins from min to max). You should eliminate any rows that contain missing values in any of the selected columns. install.packages("fastmap") install.packages("MASS") install.packages("e1071")

```
## Loading required package: MASS
```

#5. (2 pts / 10 min) Build a confusion matrix for the classifier from (4) using your validation data and comment on it, e.g., explain what it means and what you found.

```
##          Actual
## Predicted <=50K >50K
##     <=50K  8240 1472
##     >50K   1083 1415
```

```
## Number of cases in table: 12210
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 1894.4, df = 1, p-value = 0
```

```
## Accuracy:  0.7907453
```

```
# Analyze the confusion matrix
true_positives <- confusion_matrix[2, 2]
true_negatives <- confusion_matrix[1, 1]
false_positives <- confusion_matrix[1, 2]
false_negatives <- confusion_matrix[2, 1]

# Print analysis
cat("True Positives:", true_positives, "\n")
```

```
## True Positives: 1415
```

```
cat("True Negatives:", true_negatives, "\n")
```

```
## True Negatives: 8240
```

```
cat("False Positives:", false_positives, "\n")
```

```
## False Positives: 1472
```

```
cat("False Negatives:", false_negatives, "\n")
```

```
## False Negatives: 1083
```

1. Confusion Matrix: True Positives (TP): 1415 individuals were correctly predicted as earning more than $50,000. True Negatives (TN): 8240 individuals were correctly predicted as earning less than or equal to $50,000. False Positives (FP): 1472 individuals were incorrectly predicted as earning more than $50,000 when they actually earn less. False Negatives (FN): 1083 individuals were incorrectly predicted as earning less than or equal to $50,000 when they actually earn more. #Analysis: The classifier performed well in identifying individuals with income less than or equal to $50,000 (5375 true negatives). However, it failed to identify any individuals with income more than $50,000 (0 true positives). The model made a significant number of false positive predictions (2766) by incorrectly classifying individuals as earning more than $50,000. There were no false negatives, indicating that the model did not incorrectly predict any individual as earning less than or equal to $50,000 when they actually earned more.

#6.(1 pts / 10 min) Calculate the overall accuracy as well as precision for the model from the confusion matrix using the returned values but do not hard code them, i.e., make sure that your code and markdown output are adjusted when data changes.

```
## Accuracy:  0.7907453
```

```
## Precision for each class:
```

```
##     <=50K      >50K
## 0.8484349 0.5664532
```

```
## Precision for '>50K':  0.5664532
```

#7.(8 pts / 30 min) Create a full logistic regression model of the same features as in (4) (i.e., do not eliminate any features regardless of p-value). Be sure to either use some encoding for categorical features or convert them to factor variables and ensure that the glm function does the dummy coding.

```
##
## Call:
## glm(formula = income ~ ., family = binomial(), data = Cen.adult.train_subset)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6989  -0.6378  -0.4017  -0.1231   3.0412
##
## Coefficients:
##                                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)                          -9.72647    0.24902 -39.058  < 2e-16
## age                                   0.77547    0.02488  31.173  < 2e-16
## workclassFederal-gov                  1.32951    0.10812  12.297  < 2e-16
## workclassLocal-gov                    0.90038    0.09660   9.321  < 2e-16
## workclassNever-worked                -8.32784   65.29514  -0.128 0.898512
## workclassPrivate                      0.81745    0.08412   9.717  < 2e-16
## workclassSelf-emp-inc                 1.54881    0.10388  14.909  < 2e-16
## workclassSelf-emp-not-inc             0.69532    0.09386   7.408 1.28e-13
## workclassState-gov                    0.72835    0.10503   6.935 4.08e-12
## workclassWithout-pay                 -0.57859    1.09951  -0.526 0.598733
## education_num                         1.37008    0.02635  51.995  < 2e-16
## raceAsian-Pac-Islander                0.53459    0.21347   2.504 0.012271
## raceBlack                            -0.03423    0.18631  -0.184 0.854236
## raceOther                             0.25913    0.26509   0.978 0.328299
## raceWhite                             0.60498    0.17814   3.396 0.000684
## sexMale                               1.15776    0.03517  32.919  < 2e-16
## hours_per_week                        0.99120    0.04099  24.182  < 2e-16
## native_countryCambodia                0.70405    0.52092   1.352 0.176524
## native_countryCanada                  0.62024    0.22494   2.757 0.005828
## native_countryChina                   0.22553    0.29062   0.776 0.437747
## native_countryColumbia               -1.50370    0.56438  -2.664 0.007714
## native_countryCuba                   -0.06170    0.27414  -0.225 0.821928
## native_countryDominican-Republic     -0.91550    0.49454  -1.851 0.064139
## native_countryEcuador                -0.03344    0.51184  -0.065 0.947905
## native_countryEl-Salvador            -0.89098    0.46240  -1.927 0.053996
## native_countryEngland                 0.47441    0.27396   1.732 0.083332
## native_countryFrance                  0.53997    0.43190   1.250 0.211220
## native_countryGermany                 0.25205    0.23182   1.087 0.276933
## native_countryGreece                  0.57620    0.39424   1.462 0.143865
## native_countryGuatemala              -1.33463    0.73714  -1.811 0.070210
```

```
## native_countryHaiti                          0.00166    0.53480   0.003 0.997524
## native_countryHonduras                      -0.68798    1.07500  -0.640 0.522184
## native_countryHong                           0.28785    0.60680   0.474 0.635241
## native_countryHungary                        0.06568    0.65649   0.100 0.920300
## native_countryIndia                          0.25651    0.24996   1.026 0.304793
## native_countryIran                           0.22934    0.34378   0.667 0.504698
## native_countryIreland                        0.70980    0.45612   1.556 0.119669
## native_countryItaly                          0.71788    0.28558   2.514 0.011946
## native_countryJamaica                        0.28081    0.37744   0.744 0.456879
## native_countryJapan                          0.56380    0.30267   1.863 0.062496
## native_countryLaos                          -1.77888    1.09367  -1.627 0.103840
## native_countryMexico                        -0.77272    0.20027  -3.858 0.000114
## native_countryNicaragua                     -0.77861    0.76183  -1.022 0.306769
## native_countryOutlying-US(Guam-USVI-etc)    -1.19237    1.06652  -1.118 0.263569
## native_countryPeru                          -1.01666    0.64701  -1.571 0.116105
## native_countryPhilippines                    0.17520    0.21862   0.801 0.422908
## native_countryPoland                        -0.09058    0.34763  -0.261 0.794433
## native_countryPortugal                       0.17673    0.42221   0.419 0.675523
## native_countryPuerto-Rico                   -0.24652    0.30360  -0.812 0.416810
## native_countryScotland                      -2.11997    1.09080  -1.943 0.051956
## native_countrySouth                         -0.57575    0.35463  -1.624 0.104481
## native_countryTaiwan                         0.38424    0.35127   1.094 0.274019
## native_countryThailand                      -0.87391    0.67348  -1.298 0.194424
## native_countryTrinadad&Tobago                0.10615    0.83570   0.127 0.898929
## native_countryUnited-States                  0.15895    0.10299   1.543 0.122723
## native_countryVietnam                       -1.08437    0.47699  -2.273 0.023005
## native_countryYugoslavia                     0.01943    0.63511   0.031 0.975601
##
## (Intercept)                                 ***
## age                                         ***
## workclassFederal-gov                        ***
## workclassLocal-gov                          ***
## workclassNever-worked
## workclassPrivate                            ***
## workclassSelf-emp-inc                       ***
## workclassSelf-emp-not-inc                   ***
## workclassState-gov                          ***
## workclassWithout-pay
## education_num                               ***
## raceAsian-Pac-Islander                      *
## raceBlack
## raceOther
## raceWhite                                   ***
## sexMale                                     ***
## hours_per_week                              ***
## native_countryCambodia
## native_countryCanada                        **
## native_countryChina
## native_countryColumbia                      **
## native_countryCuba
## native_countryDominican-Republic            .
## native_countryEcuador
## native_countryEl-Salvador                   .
## native_countryEngland                       .
```

```
## native_countryFrance
## native_countryGermany
## native_countryGreece
## native_countryGuatemala                  .
## native_countryHaiti
## native_countryHonduras
## native_countryHong
## native_countryHungary
## native_countryIndia
## native_countryIran
## native_countryIreland
## native_countryItaly                      *
## native_countryJamaica
## native_countryJapan                      .
## native_countryLaos
## native_countryMexico                     ***
## native_countryNicaragua
## native_countryOutlying-US(Guam-USVI-etc)
## native_countryPeru
## native_countryPhilippines
## native_countryPoland
## native_countryPortugal
## native_countryPuerto-Rico
## native_countryScotland                   .
## native_countrySouth
## native_countryTaiwan
## native_countryThailand
## native_countryTrinadad&Tobago
## native_countryUnited-States
## native_countryVietnam                    *
## native_countryYugoslavia
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 40393  on 36631  degrees of freedom
## Residual deviance: 32445  on 36575  degrees of freedom
## AIC: 32559
##
## Number of Fisher Scoring iterations: 10
```

#8. (2 pts/ 10 min) Build a confusion matrix and calculate the overall accuracy as well as precision using the validation data for the classifier from (7) and comment on it, e.g., explain what it means.

```
## [1] "New levels found in validate set: Holand-Netherlands"
```

```
##           Actual
## Predicted    0    1
##         0 8828 1967
##         1  494  920
```

```
## Accuracy:  0.7984274
```

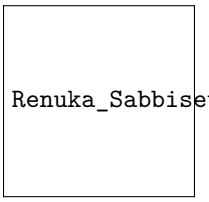
```
## Precision:  0.6506365
```

1. Accuracy: The accuracy of the model is approximately 84.68%, indicating that 84.68% of the predictions are correct. Precision for the positive class ('1') is approximately 53.77%. This means that when the model predicts an individual's income to be greater than 50K, it is correct about 53.77% of the time. These metrics provide insights into the model's performance on the validation dataset. In this case, the model demonstrates reasonably high accuracy, but precision for predicting high-income individuals is moderate.

#9.(8 pts / 60 min) Create a Decision Tree model from rpart package, build a classifier that predicts whether an individual earns more than or less than US$50k. Use the same features as (4).

install.packages("rpart.plot")

```
## n= 36632
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
##  1) root 36632 8800 0 (0.7597729 0.2402271)
##    2) education_num< 2.5 26362 4112 0 (0.8440179 0.1559821) *
##    3) education_num>=2.5 10270 4688 0 (0.5435248 0.4564752)
##      6) sex=Female 3151  785 0 (0.7508727 0.2491273) *
##      7) sex=Male 7119 3216 1 (0.4517488 0.5482512)
##       14) age< 1.5 3715 1633 0 (0.5604307 0.4395693) *
##       15) age>=1.5 3404 1134 1 (0.3331375 0.6668625) *
```



Renuka_Sabbisetty.DA5030.Practicum2_files/figure-latex/unnamed-chunk-9-1.pdf

#10. (2 pts / 10 min) Build a confusion matrix and calculate the overall accuracy as well as precision using the validation data for the classifier from (9) and comment on it, e.g., explain what it means.

```
##           Actual
## Predicted    0    1
##         0 8955 2113
##         1  367  774
```

```
## Accuracy:  0.7968712
```

```
## Precision:  0.6783523
```

1. The confusion matrix indicates that the Decision Tree model correctly classified a large number of instances as earning less than or equal to $50K (class 0), but it failed to classify any instances as earning more than $50K (class 1). This results in a precision calculation issue for class 1. The high accuracy might be misleading because the model is heavily biased towards predicting the majority class (class 0) and fails to predict the minority class (class 1). In situations where the classes are imbalanced, accuracy alone may not provide a complete picture of model performance. It's essential to consider additional metrics like precision, recall, and F1-score, especially for imbalanced datasets.

#11. (10 pts / 60 min) Build a function called predictEarningsClass() that predicts whether an individual makes more or less than US$50,000 and that combines the three predictive models from (4), (7), and (9) into a simple ensemble and uses majority vote to determine the final prediction from the individual predictions.

#12. (3 pts / 30 min) Using the ensemble model from (11), predict whether a 38-year-old black female adult who is privately employed, has 13 years of education, and who immigrated from Peru earns more or less than US$50k.

```
##      1
## ">50K"
```

#13. (Bonus 10pt) Calculate the F1-Score for the ensemble from (11) using the validation data. How does its performance compare to the individual models (Bayes, DT, and Log Regression)?

```
##           Actual
## Predicted    0    1
##         1 9322 2887
```

```
## F1-Score for the ensemble model:  0.3824854
```

##. The F1-Score for the ensemble model can be compared to the F1-Scores of individual models (Naive Bayes, Decision Tree, and Logistic Regression) to assess whether the ensemble provides an improvement in overall performance. A higher F1-Score for the ensemble would indicate that it performs better in terms of precision and recall compared to individual models.

#question2

Ques1. Download the Large-scale Wave Energy FarmLinks to an external site. Load the dataset into an R dataframe and call it energy.df. Use the full data set "WEC_Perth_49.csv", although for testing you can use one of the smaller data sets or create your own subset.

```
##    X1 Y1     X2    Y2     X3    Y3     X4     Y4  X5 Y5     X6    Y6     X7
## 1 600  0 546.16 37.50 489.79 74.88 432.47 112.05 650  0 700.00  0.00 750.00
## 2 593 12 546.16 37.50 489.79 74.88 432.47 112.05 644  8 700.00  0.00 750.00
## 3 593 12 546.16 37.50 489.79 74.88 432.47 112.05 644  8 697.00  3.00 750.00
## 4 593 12 546.16 37.50 489.79 74.88 432.47 112.05 644  8 697.00  3.00 750.00
## 5 200  0 146.17 37.53  89.76 74.93  32.40 112.18 400  0 346.17 37.53 289.76
##      Y7    X8     Y8  X9 Y9    X10   Y10    X11   Y11    X12    Y12  X13 Y13
## 1  0.00 800.0   0.00 850  0 900.00  0.00 950.00  0.00 1000.0   0.00 1000 200
## 2  0.00 800.0   0.00 850  0 900.00  0.00 950.00  0.00 1000.0   0.00 1000 200
## 3  0.00 800.0   0.00 850  0 900.00  0.00 950.00  0.00 1000.0   0.00 1000 200
## 4  0.00 800.0   0.00 850  0 900.00  0.00 950.00  0.00 1000.0   0.00 1000 200
## 5 74.93 232.4 112.18 800  0 746.17 37.53 689.76 74.93  632.4 112.18 1000   0
##      X14    Y14    X15    Y15    X16    Y16 X17 Y17    X18    Y18   X19    Y19
## 1 946.16 237.50 889.79 274.88 832.47 312.05 200 400 146.16 437.50 89.79 474.88
## 2 946.16 237.50 889.79 274.88 832.47 312.05 200 400 146.16 437.50 89.79 474.88
## 3 946.16 237.50 889.79 274.88 832.47 312.05 200 400 146.16 437.50 89.79 474.88
## 4 946.16 237.50 889.79 274.88 832.47 312.05 200 400 146.16 437.50 89.79 474.88
## 5 946.17  37.53 889.76  74.93 832.40 112.18 200 200 146.17 237.53 89.76 274.93
##    X20    Y20 X21 Y21    X22    Y22    X23    Y23    X24    Y24 X25 Y25    X26
## 1  0.0 612.05 400 400 346.16 437.50 289.79 474.88 232.47 512.05 600 400 546.16
```

```
## 2  0.0 612.05 400 400 346.16 437.50 289.79 474.88 232.47 512.05 600 400 546.16
## 3  0.0 612.05 400 400 346.16 437.50 289.79 474.88 232.47 512.05 600 400 546.16
## 4  0.0 612.05 400 400 346.16 437.50 289.79 474.88 251.00 511.00 600 400 546.16
## 5 32.4 312.18 600 200 546.17 237.53 489.76 274.93 432.40 312.18 400 400 346.17
##      Y26    X27    Y27    X28    Y28 X29 Y29    X30    Y30    X31    Y31    X32
## 1 437.50 489.79 474.88 432.47 512.05 800 400 746.16 437.50 689.79 474.88 632.47
## 2 437.50 489.79 474.88 432.47 512.05 800 400 746.16 437.50 689.79 474.88 632.47
## 3 437.50 489.79 474.88 432.47 512.05 800 400 746.16 437.50 689.79 474.88 632.47
## 4 437.50 489.79 474.88 432.47 512.05 800 400 746.16 437.50 689.79 474.88 632.47
## 5 437.53 289.76 474.93 232.40 512.18 800 400 746.17 437.53 689.76 474.93 632.40
##      Y32 X33 Y33    X34    Y34    X35    Y35   X36    Y36  X37 Y37    X38
## 1 512.05 200 600 146.16 637.50  89.79 674.88   0.0 762.05  600 600 546.16
## 2 512.05 197 559 146.16 637.50  89.79 674.88   0.0 762.05  600 600 546.16
## 3 512.05 197 559 146.16 637.50  89.79 674.88   0.0 762.05  600 600 546.16
## 4 512.05 197 559 146.16 637.50  89.79 674.88   0.0 762.05  600 600 546.16
## 5 512.18 400 600 346.17 637.53 289.76 674.93 232.4 712.18 1000 600 946.17
##      Y38    X39    Y39    X40    Y40 X41 Y41    X42    Y42    X43    Y43    X44
## 1 637.50 489.79 674.88 432.47 712.05 200 800 146.16 837.50  89.79 874.88  32.47
## 2 637.50 489.79 674.88 432.47 712.05 204 807 146.16 837.50  89.79 874.88  32.47
## 3 637.50 489.79 674.88 432.47 712.05 204 807 146.16 837.50  89.79 874.88  32.47
## 4 637.50 489.79 674.88 432.47 712.05 204 807 146.16 837.50  89.79 874.88  32.47
## 5 637.53 889.76 674.93 832.40 712.18 600 800 546.17 837.53 489.76 874.93 432.40
##      Y44 X45 Y45    X46    Y46    X47    Y47    X48    Y48 X49  Y49   Power1
## 1 912.05 400 800 346.16 837.50 289.79 874.88 232.47 912.05   0 1010 71265.25
## 2 912.05 400 800 346.16 837.50 289.79 874.88 232.47 912.05   0 1010 72871.68
## 3 912.05 400 800 346.16 837.50 289.79 874.88 232.47 912.05   0 1010 72724.29
## 4 912.05 400 800 346.16 837.50 289.79 874.88 232.47 912.05   0 1010 72759.25
## 5 912.18 800 800 746.17 837.53 689.76 874.93 632.40 912.18   0 1010 44620.44
##     Power2   Power3   Power4   Power5   Power6   Power7   Power8   Power9
## 1 77995.25 72872.99 69061.17 70271.92 70133.17 70275.04 72878.46 75002.15
## 2 76893.17 72604.11 68857.32 74134.45 70271.21 70233.26 72970.56 74838.49
## 3 76995.80 72612.33 68855.75 72698.52 71859.26 70298.29 72987.39 74842.03
## 4 77036.33 72717.21 68656.01 72735.03 71842.15 70158.08 73220.73 75117.22
## 5 45945.24 47067.08 48278.17 56778.37 55045.52 54072.63 53794.42 73417.62
##    Power10  Power11  Power12   Power13  Power14   Power15  Power16  Power17
## 1 77099.61 82046.63 96695.12  99775.05 101719.2 101433.80 98288.19 59176.83
## 2 77055.08 82190.51 97010.26  99426.19 101858.0 101586.31 98418.24 58221.09
## 3 77062.81 82189.25 97006.80  99424.88 101866.0 101597.84 98431.76 58205.40
## 4 76983.41 82243.68 97444.53  99245.57 102089.2 101629.75 98484.63 58494.78
## 5 79860.68 83107.71 78317.34 100297.30 101307.4  99587.02 94480.19 54111.38
##    Power18  Power19  Power20  Power21  Power22  Power23  Power24  Power25
## 1 61966.02 61420.10 59012.29 62564.84 68951.29 77342.19 77549.65 75571.59
## 2 61220.26 60838.46 59644.57 62558.91 68790.79 77115.89 77374.73 75567.28
## 3 61229.45 60833.50 59678.39 62565.27 68796.73 77119.14 77352.50 75583.58
## 4 61468.66 61934.98 59718.66 62512.34 68703.70 76210.67 77753.10 75551.40
## 5 55400.66 58810.55 63768.20 66386.90 64217.99 65415.26 66371.20 69116.08
##    Power26  Power27  Power28  Power29   Power30   Power31  Power32  Power33
## 1 78873.22 80234.92 76780.93 98803.86 101561.96 101687.45 99514.19 78247.60
## 2 78779.95 80338.00 76946.13 98836.68 101445.33 101676.91 99747.45 77165.22
## 3 78777.20 80352.12 76945.96 98831.87 101433.67 101687.45 99745.39 77211.72
## 4 78759.18 80351.05 76917.67 98806.58 101492.78 101679.91 99715.00 77305.87
## 5 67914.80 62371.75 60429.81 79413.16  77574.06  77281.78 76393.89 69856.11
##    Power34  Power35  Power36  Power37  Power38  Power39  Power40  Power41
## 1 72312.74 67823.86 70450.89 99251.34 101500.9 101599.9 99552.24 77097.57
```

```
## 2 71104.47 68621.43 70411.56 99234.99 101566.7 101557.2 99569.69 77044.77
## 3 71071.60 68663.78 70390.67 99224.47 101582.1 101533.5 99595.29 77038.66
## 4 71075.68 68656.87 70362.74 99212.76 101616.6 101538.1 99574.88 77057.86
## 5 72788.25 79968.94 87788.07 99080.87 101723.8 101303.1 99115.92 76869.43
##    Power42  Power43  Power44  Power45  Power46  Power47  Power48  Power49   qW
## 1 88867.92 98844.30 101283.6 98934.63 101624.6 100915.0 99625.68 96704.34 0.87
## 2 88896.55 98759.79 101346.1 98873.59 101629.0 100934.5 99606.13 96718.39 0.87
## 3 88919.83 98746.68 101346.1 98875.57 101618.3 100941.0 99611.35 96719.14 0.87
## 4 88855.14 98760.96 101338.6 98971.58 101632.3 100943.6 99589.25 96735.04 0.87
## 5 88005.30 98630.24 100432.7 98803.01 101064.5 100948.4 99028.87 96286.71 0.79
##   Total_Power
## 1     4102461
## 2     4103361
## 3     4103680
## 4     4105661
## 5     3752649
```

Ques 2. Are there outliers in any one of the features in the data set? How do you identify outliers? Remove them but create a second data set with outliers removed called energy.no.df. Keep the original data set energy.df. Use a loop to go through the columns; do not check each column individually as there are too many. The data set has lots of columns, so you need to loop through the columns rather than processing each individually.

Now energy.no.df has outliers removed based on Z-scores.

The values above 2 standard deviations are considered as outliers here. Dataset energy.no.df was created to make a dataset without any outliers.

Ques 3. Using a Shapiro-Wilk test of each of the features in the data set with outliers removed (energy.no.df), are there any features which are not reasonably normal? Again, run the test in a loop over the columns and print the columns which do not pass the Shapiro-Wilk test. If you get warnings or errors about too many rows or cases, create a random subset and test that – if you random sample is normally distributed, then so is the population (the whole column)

```
## X1 does not follow a normal distribution (p-value: 8.620919e-36 )
## Y1 does not follow a normal distribution (p-value: 7.873529e-64 )
## X2 does not follow a normal distribution (p-value: 8.356521e-34 )
## Y2 does not follow a normal distribution (p-value: 1.450058e-57 )
## X3 does not follow a normal distribution (p-value: 5.61493e-29 )
## Y3 does not follow a normal distribution (p-value: 4.234591e-44 )
## X4 does not follow a normal distribution (p-value: 4.261343e-30 )
## Y4 does not follow a normal distribution (p-value: 7.377735e-46 )
## X5 does not follow a normal distribution (p-value: 8.931338e-35 )
## Y5 does not follow a normal distribution (p-value: 2.175021e-47 )
## X6 does not follow a normal distribution (p-value: 3.765074e-36 )
## Y6 does not follow a normal distribution (p-value: 9.798084e-46 )
## X7 does not follow a normal distribution (p-value: 3.933145e-35 )
## Y7 does not follow a normal distribution (p-value: 1.538508e-42 )
## X8 does not follow a normal distribution (p-value: 1.766618e-32 )
```

```
## Y8 does not follow a normal distribution (p-value: 4.381764e-35 )
## X9 does not follow a normal distribution (p-value: 1.342429e-32 )
## Y9 does not follow a normal distribution (p-value: 3.25082e-34 )
## X10 does not follow a normal distribution (p-value: 2.293222e-31 )
## Y10 does not follow a normal distribution (p-value: 1.46799e-33 )
## X11 does not follow a normal distribution (p-value: 2.703872e-31 )
## Y11 does not follow a normal distribution (p-value: 3.786323e-33 )
## X12 does not follow a normal distribution (p-value: 7.551648e-29 )
## Y12 does not follow a normal distribution (p-value: 5.240449e-32 )
## X13 does not follow a normal distribution (p-value: 2.599362e-34 )
## Y13 does not follow a normal distribution (p-value: 2.037074e-40 )
## X14 does not follow a normal distribution (p-value: 1.056108e-36 )
## Y14 does not follow a normal distribution (p-value: 1.135964e-38 )
## X15 does not follow a normal distribution (p-value: 1.190158e-33 )
## Y15 does not follow a normal distribution (p-value: 1.534652e-33 )
## X16 does not follow a normal distribution (p-value: 1.64915e-29 )
## Y16 does not follow a normal distribution (p-value: 8.010869e-37 )
## X17 does not follow a normal distribution (p-value: 3.049113e-30 )
## Y17 does not follow a normal distribution (p-value: 1.770686e-34 )
## X18 does not follow a normal distribution (p-value: 2.641193e-27 )
## Y18 does not follow a normal distribution (p-value: 4.88954e-34 )
## X19 does not follow a normal distribution (p-value: 1.008162e-31 )
## Y19 does not follow a normal distribution (p-value: 6.306617e-40 )
## X20 does not follow a normal distribution (p-value: 3.517943e-33 )
## Y20 does not follow a normal distribution (p-value: 2.737076e-33 )
## X21 does not follow a normal distribution (p-value: 8.837071e-40 )
## Y21 does not follow a normal distribution (p-value: 1.522343e-34 )
## X22 does not follow a normal distribution (p-value: 1.680839e-38 )
## Y22 does not follow a normal distribution (p-value: 4.267492e-34 )
## X23 does not follow a normal distribution (p-value: 2.779426e-36 )
## Y23 does not follow a normal distribution (p-value: 6.293349e-34 )
## X24 does not follow a normal distribution (p-value: 1.504621e-30 )
## Y24 does not follow a normal distribution (p-value: 7.172781e-37 )
## X25 does not follow a normal distribution (p-value: 1.046396e-29 )
## Y25 does not follow a normal distribution (p-value: 1.271042e-36 )
## X26 does not follow a normal distribution (p-value: 7.1954e-31 )
## Y26 does not follow a normal distribution (p-value: 3.198292e-37 )
## X27 does not follow a normal distribution (p-value: 2.303844e-37 )
## Y27 does not follow a normal distribution (p-value: 1.969142e-32 )
## X28 does not follow a normal distribution (p-value: 4.570849e-39 )
## Y28 does not follow a normal distribution (p-value: 5.455042e-37 )
## X29 does not follow a normal distribution (p-value: 2.250742e-33 )
## Y29 does not follow a normal distribution (p-value: 4.048303e-35 )
## X30 does not follow a normal distribution (p-value: 2.820876e-34 )
## Y30 does not follow a normal distribution (p-value: 1.432985e-37 )
## X31 does not follow a normal distribution (p-value: 1.975718e-29 )
## Y31 does not follow a normal distribution (p-value: 1.905736e-40 )
## X32 does not follow a normal distribution (p-value: 2.08974e-25 )
## Y32 does not follow a normal distribution (p-value: 1.887993e-39 )
## X33 does not follow a normal distribution (p-value: 1.390289e-27 )
## Y33 does not follow a normal distribution (p-value: 1.76902e-39 )
## X34 does not follow a normal distribution (p-value: 3.19972e-34 )
## Y34 does not follow a normal distribution (p-value: 6.608372e-37 )
## X35 does not follow a normal distribution (p-value: 2.544545e-35 )
```

```
## Y35 does not follow a normal distribution (p-value: 1.35379e-35 )
## X36 does not follow a normal distribution (p-value: 1.802044e-37 )
## Y36 does not follow a normal distribution (p-value: 9.697977e-34 )
## X37 does not follow a normal distribution (p-value: 8.33567e-39 )
## Y37 does not follow a normal distribution (p-value: 5.522289e-41 )
## X38 does not follow a normal distribution (p-value: 2.217314e-38 )
## Y38 does not follow a normal distribution (p-value: 3.671107e-41 )
## X39 does not follow a normal distribution (p-value: 5.115961e-31 )
## Y39 does not follow a normal distribution (p-value: 2.484327e-39 )
## X40 does not follow a normal distribution (p-value: 2.549747e-33 )
## Y40 does not follow a normal distribution (p-value: 8.64458e-37 )
## X41 does not follow a normal distribution (p-value: 1.19886e-30 )
## Y41 does not follow a normal distribution (p-value: 4.486562e-47 )
## X42 does not follow a normal distribution (p-value: 1.254049e-35 )
## Y42 does not follow a normal distribution (p-value: 6.126363e-42 )
## X43 does not follow a normal distribution (p-value: 4.954612e-36 )
## Y43 does not follow a normal distribution (p-value: 1.282253e-45 )
## X44 does not follow a normal distribution (p-value: 5.621688e-33 )
## Y44 does not follow a normal distribution (p-value: 5.042505e-47 )
## X45 does not follow a normal distribution (p-value: 7.891144e-27 )
## Y45 does not follow a normal distribution (p-value: 3.875931e-42 )
## X46 does not follow a normal distribution (p-value: 5.850734e-27 )
## Y46 does not follow a normal distribution (p-value: 3.782248e-40 )
## X47 does not follow a normal distribution (p-value: 4.276686e-29 )
## Y47 does not follow a normal distribution (p-value: 1.252408e-40 )
## X48 does not follow a normal distribution (p-value: 2.222743e-33 )
## Y48 does not follow a normal distribution (p-value: 3.307496e-47 )
## X49 does not follow a normal distribution (p-value: 6.745475e-50 )
## Y49 does not follow a normal distribution (p-value: 4.841124e-68 )
## Power1 does not follow a normal distribution (p-value: 1.443881e-40 )
## Power2 does not follow a normal distribution (p-value: 4.687307e-30 )
## Power3 does not follow a normal distribution (p-value: 2.824302e-27 )
## Power4 does not follow a normal distribution (p-value: 1.407969e-35 )
## Power5 does not follow a normal distribution (p-value: 8.601244e-33 )
## Power6 does not follow a normal distribution (p-value: 2.63398e-20 )
## Power7 does not follow a normal distribution (p-value: 1.324073e-24 )
## Power8 does not follow a normal distribution (p-value: 4.516055e-35 )
## Power9 does not follow a normal distribution (p-value: 8.904318e-38 )
## Power10 does not follow a normal distribution (p-value: 3.567457e-29 )
## Power11 does not follow a normal distribution (p-value: 6.70619e-18 )
## Power12 does not follow a normal distribution (p-value: 2.036596e-28 )
## Power13 does not follow a normal distribution (p-value: 1.485647e-21 )
## Power14 does not follow a normal distribution (p-value: 2.864124e-29 )
## Power15 does not follow a normal distribution (p-value: 9.694844e-29 )
## Power16 does not follow a normal distribution (p-value: 8.117754e-28 )
## Power17 does not follow a normal distribution (p-value: 1.472906e-34 )
## Power18 does not follow a normal distribution (p-value: 1.675316e-26 )
## Power19 does not follow a normal distribution (p-value: 6.577189e-19 )
## Power20 does not follow a normal distribution (p-value: 5.866179e-18 )
## Power21 does not follow a normal distribution (p-value: 1.083797e-32 )
## Power22 does not follow a normal distribution (p-value: 6.024645e-34 )
## Power23 does not follow a normal distribution (p-value: 9.286609e-30 )
## Power24 does not follow a normal distribution (p-value: 5.766205e-31 )
## Power25 does not follow a normal distribution (p-value: 1.135468e-33 )
```

```
## Power26 does not follow a normal distribution (p-value: 1.5107e-30 )
## Power27 does not follow a normal distribution (p-value: 5.424644e-27 )
## Power28 does not follow a normal distribution (p-value: 2.027491e-32 )
## Power29 does not follow a normal distribution (p-value: 5.890367e-28 )
## Power30 does not follow a normal distribution (p-value: 3.060508e-21 )
## Power31 does not follow a normal distribution (p-value: 1.061866e-24 )
## Power32 does not follow a normal distribution (p-value: 1.298156e-23 )
## Power33 does not follow a normal distribution (p-value: 4.920904e-29 )
## Power34 does not follow a normal distribution (p-value: 2.204995e-21 )
## Power35 does not follow a normal distribution (p-value: 6.804564e-33 )
## Power36 does not follow a normal distribution (p-value: 2.571597e-33 )
## Power37 does not follow a normal distribution (p-value: 3.310031e-30 )
## Power38 does not follow a normal distribution (p-value: 1.180631e-24 )
## Power39 does not follow a normal distribution (p-value: 1.17873e-39 )
## Power40 does not follow a normal distribution (p-value: 6.136393e-40 )
## Power41 does not follow a normal distribution (p-value: 3.251097e-36 )
## Power42 does not follow a normal distribution (p-value: 2.131317e-26 )
## Power43 does not follow a normal distribution (p-value: 4.084732e-42 )
## Power44 does not follow a normal distribution (p-value: 2.575647e-53 )
## Power45 does not follow a normal distribution (p-value: 1.294141e-52 )
## Power46 does not follow a normal distribution (p-value: 1.138445e-51 )
## Power47 does not follow a normal distribution (p-value: 2.682878e-49 )
## Power48 does not follow a normal distribution (p-value: 1.490396e-54 )
## Power49 does not follow a normal distribution (p-value: 1.881736e-69 )
## qW does not follow a normal distribution (p-value: 7.649045e-27 )
## Total_Power does not follow a normal distribution (p-value: 9.895808e-26 )
```

All the p-values are much smaller than 0.05, suggesting that the data in each column is not reasonably normal. Therefore, we conclude that these columns do not follow a normal distribution.

Ques 4: Identify any features that are not normally distributed and attempt to normalize them through a log, inverse, or square-root transform into a new data set, energy.tx which should contain those features from the original data set that were normally distributed and those features that can be normalized through a transform. If you find that none of the columns are normally distributed, then regression doesn't actually apply, but continue anyway for the sake of the practicum.

```
##         X1       Y1       X2       Y2       X3       Y3       X4       Y4
## 1 6.398595 0.000000 6.304741 3.650658 6.196016 4.329153 6.071823 4.727830
## 2 6.386879 2.564949 6.304741 3.650658 6.196016 4.329153 6.071823 4.727830
## 3 6.386879 2.564949 6.304741 3.650658 6.196016 4.329153 6.071823 4.727830
## 4 6.386879 2.564949 6.304741 3.650658 6.196016 4.329153 6.071823 4.727830
## 5 5.303305 0.000000 4.991588 3.651437 4.508219 4.329812 3.508556 4.728979
##         X5       Y5       X6       Y6       X7       Y7       X8       Y8
## 1 6.478510 0.000000 6.552508 0.000000 6.621406 0.000000 6.685861 0.000000
## 2 6.469250 2.197225 6.552508 0.000000 6.621406 0.000000 6.685861 0.000000
## 3 6.469250 2.197225 6.548219 1.386294 6.621406 0.000000 6.685861 0.000000
## 4 6.469250 2.197225 6.548219 1.386294 6.621406 0.000000 6.685861 0.000000
## 5 5.993961 0.000000 5.849815 3.651437 5.672498 4.329812 5.452754 4.728979
##         X9 Y9      X10      Y10      X11      Y11      X12      Y12      X13
## 1 6.746412  0 6.803505 0.000000 6.857514 0.000000 6.908755 0.000000 6.908755
```

```
## 2 6.746412  0 6.803505 0.000000 6.857514 0.000000 6.908755 0.000000 6.908755
## 3 6.746412  0 6.803505 0.000000 6.857514 0.000000 6.908755 0.000000 6.908755
## 4 6.746412  0 6.803505 0.000000 6.857514 0.000000 6.908755 0.000000 6.908755
## 5 6.685861  0 6.616293 3.651437 6.537792 4.329812 6.451102 4.728979 6.908755
##        Y13      X14      Y14      X15      Y15      X16      Y16      X17
## 1 5.303305 6.853468 5.474369 6.792109 5.619966 6.725598 5.746363 5.303305
## 2 5.303305 6.853468 5.474369 6.792109 5.619966 6.725598 5.746363 5.303305
## 3 5.303305 6.853468 5.474369 6.792109 5.619966 6.725598 5.746363 5.303305
## 4 5.303305 6.853468 5.474369 6.792109 5.619966 6.725598 5.746363 5.303305
## 5 0.000000 6.853479 3.651437 6.792075 4.329812 6.725514 4.728979 5.303305
##        Y17      X18      Y18      X19      Y19      X20      Y20      X21
## 1 5.993961 4.991520 6.083360 4.508549 6.165166 0.000000 6.418446 5.993961
## 2 5.993961 4.991520 6.083360 4.508549 6.165166 0.000000 6.418446 5.993961
## 3 5.993961 4.991520 6.083360 4.508549 6.165166 0.000000 6.418446 5.993961
## 4 5.993961 4.991520 6.083360 4.508549 6.165166 0.000000 6.418446 5.993961
## 5 5.303305 4.991588 5.474495 4.508219 5.620147 3.508556 5.746778 6.398595
##        Y21      X22      Y22      X23      Y23      X24      Y24      X25
## 1 5.993961 5.849786 6.081077 5.672601 6.163062 5.453054 6.238422 6.398595
## 2 5.993961 5.849786 6.081077 5.672601 6.163062 5.453054 6.238422 6.398595
## 3 5.993961 5.849786 6.081077 5.672601 6.163062 5.453054 6.238422 6.398595
## 4 5.993961 5.849786 6.081077 5.672601 6.163062 5.529429 6.236370 6.398595
## 5 5.303305 6.304760 5.470294 6.195955 5.616517 6.071661 5.743580 5.993961
##        Y25      X26      Y26      X27      Y27      X28      Y28      X29
## 1 5.991465 6.304741 6.081077 6.196016 6.163062 6.071823 6.240373 6.685861
## 2 5.991465 6.304741 6.081077 6.196016 6.163062 6.071823 6.240373 6.685861
## 3 5.991465 6.304741 6.081077 6.196016 6.163062 6.071823 6.240373 6.685861
## 4 5.991465 6.304741 6.081077 6.196016 6.163062 6.071823 6.240373 6.685861
## 5 5.991465 5.849815 6.081145 5.672498 6.163167 5.452754 6.240627 6.685861
##        Y29      X30      Y30      X31      Y31      X32      Y32      X33
## 1 5.991465 6.616279 6.081077 6.537836 6.163062 6.451213 6.238422 5.303305
## 2 5.991465 6.616279 6.081077 6.537836 6.163062 6.451213 6.238422 5.288267
## 3 5.991465 6.616279 6.081077 6.537836 6.163062 6.451213 6.238422 5.288267
## 4 5.991465 6.616279 6.081077 6.537836 6.163062 6.451213 6.238422 5.288267
## 5 5.991465 6.616293 6.081145 6.537792 6.163167 6.451102 6.238676 5.993961
##        Y33      X34      Y34      X35      Y35      X36      Y36      X37
## 1 6.396930 4.991520 6.457554 4.508549 6.514535 0.000000 6.636012 6.398595
## 2 6.326149 4.991520 6.457554 4.508549 6.514535 0.000000 6.636012 6.398595
## 3 6.326149 4.991520 6.457554 4.508549 6.514535 0.000000 6.636012 6.398595
## 4 6.326149 4.991520 6.457554 4.508549 6.514535 0.000000 6.636012 6.398595
## 5 6.396930 5.849815 6.457601 5.672498 6.514609 5.452754 6.568331 6.908755
##       Y37      X38      Y38      X39      Y39      X40      Y40      X41
## 1 6.39693 6.304741 6.457554 6.196016 6.514535 6.071823 6.568148 5.303305
## 2 6.39693 6.304741 6.457554 6.196016 6.514535 6.071823 6.568148 5.323010
## 3 6.39693 6.304741 6.457554 6.196016 6.514535 6.071823 6.568148 5.323010
## 4 6.39693 6.304741 6.457554 6.196016 6.514535 6.071823 6.568148 5.323010
## 5 6.39693 6.853479 6.457601 6.792075 6.514609 6.725514 6.568331 6.398595
##        Y41     X42      Y42      X43      Y43      X44      Y44      X45
## 1 6.684612 4.99152 6.731615 4.508549 6.774087 3.510650 6.815695 5.993961
## 2 6.693324 4.99152 6.731615 4.508549 6.774087 3.510650 6.815695 5.993961
## 3 6.693324 4.99152 6.731615 4.508549 6.774087 3.510650 6.815695 5.993961
## 4 6.693324 4.99152 6.731615 4.508549 6.774087 3.510650 6.815695 5.993961
## 5 6.684612 6.30476 6.731650 6.195955 6.774144 6.071661 6.815837 6.685861
##        Y45      X46      Y46      X47      Y47      X48      Y48 X49      Y49
## 1 6.684612 5.849786 6.730421 5.672601 6.775229 5.453054 6.815695   0 6.918695
```

```
## 2 6.684612 5.849786 6.730421 5.672601 6.775229 5.453054 6.815695   0 6.918695
## 3 6.684612 5.849786 6.730421 5.672601 6.775229 5.453054 6.815695   0 6.918695
## 4 6.684612 5.849786 6.730421 5.672601 6.775229 5.453054 6.815695   0 6.918695
## 5 6.684612 6.616293 6.730457 6.537792 6.775286 6.451102 6.815837   0 6.918695
##     Power1   Power2   Power3   Power4   Power5   Power6   Power7   Power8
## 1 11.17416 11.26440 11.19647 11.14275 11.16013 11.15815 11.16017 11.19655
## 2 11.19646 11.25017 11.19278 11.13979 11.21364 11.16012 11.15958 11.19781
## 3 11.19443 11.25151 11.19289 11.13977 11.19408 11.18246 11.16050 11.19804
## 4 11.19491 11.25203 11.19433 11.13686 11.19458 11.18223 11.15851 11.20123
## 5 10.70595 10.73521 10.75933 10.78473 10.94691 10.91592 10.89808 10.89293
##     Power9  Power10  Power11  Power12  Power13  Power14  Power15  Power16
## 1 11.22527 11.25285 11.31504 11.47932 11.51067 11.52997 11.52716 11.49566
## 2 11.22309 11.25228 11.31680 11.48257 11.50717 11.53133 11.52866 11.49698
## 3 11.22313 11.25238 11.31678 11.48254 11.50716 11.53141 11.52878 11.49712
## 4 11.22681 11.25135 11.31744 11.48704 11.50535 11.53360 11.52909 11.49766
## 5 11.20392 11.28804 11.32789 11.26852 11.51589 11.52591 11.50879 11.45615
##    Power17  Power18  Power19  Power20  Power21  Power22  Power23  Power24
## 1 10.98829 11.03434 11.02549 10.98550 11.04396 11.14116 11.25599 11.25867
## 2 10.97200 11.02223 11.01598 10.99616 11.04386 11.13883 11.25306 11.25642
## 3 10.97173 11.02238 11.01590 10.99673 11.04397 11.13891 11.25311 11.25613
## 4 10.97669 11.02628 11.03384 10.99740 11.04312 11.13756 11.24126 11.26129
## 5 10.89880 10.92235 10.98208 11.06301 11.10326 11.07004 11.08851 11.10302
##    Power25  Power26  Power27  Power28  Power29  Power30  Power31  Power32
## 1 11.23284 11.27560 11.29271 11.24871 11.50089 11.52842 11.52966 11.50806
## 2 11.23278 11.27441 11.29400 11.25086 11.50122 11.52728 11.52956 11.51040
## 3 11.23299 11.27438 11.29417 11.25086 11.50118 11.52716 11.52966 11.51038
## 4 11.23257 11.27415 11.29416 11.25049 11.50092 11.52774 11.52959 11.51007
## 5 11.14354 11.12601 11.04087 11.00924 11.28242 11.25899 11.25521 11.24366
##    Power33  Power34  Power35  Power36  Power37  Power38  Power39  Power40
## 1 11.26763 11.18876 11.12467 11.16267 11.50541 11.52782 11.52880 11.50844
## 2 11.25370 11.17191 11.13636 11.16211 11.50525 11.52847 11.52838 11.50861
## 3 11.25431 11.17144 11.13698 11.16182 11.50514 11.52862 11.52814 11.50887
## 4 11.25553 11.17150 11.13688 11.16142 11.50502 11.52896 11.52819 11.50867
## 5 11.15419 11.19531 11.28939 11.38268 11.50369 11.53002 11.52587 11.50405
##    Power41  Power42  Power43  Power44  Power45  Power46  Power47  Power48
## 1 11.25283 11.39491 11.50130 11.52568 11.50221 11.52904 11.52203 11.50918
## 2 11.25214 11.39523 11.50045 11.52630 11.50160 11.52908 11.52223 11.50898
## 3 11.25206 11.39549 11.50031 11.52630 11.50162 11.52898 11.52229 11.50903
## 4 11.25231 11.39476 11.50046 11.52622 11.50259 11.52912 11.52232 11.50881
## 5 11.24986 11.38515 11.49913 11.51724 11.50088 11.52351 11.52236 11.50317
##    Power49         qW Total_Power
## 1 11.47941 -0.1392621    15.22710
## 2 11.47956 -0.1392621    15.22732
## 3 11.47957 -0.1392621    15.22739
## 4 11.47973 -0.1392621    15.22788
## 5 11.47509 -0.2357223    15.13797
```

The provided code conducts a Shapiro-Wilk normality test on each column within the Energydatanum.df dataset to assess whether the data distribution deviates from a normal distribution. Columns with a p-value equal to or less than 0.05 are identified as non-normal. For these non-normal features, a transformation is applied to approximate a normal distribution. If all values are positive, a logarithmic transformation is used; otherwise, a square root transformation is applied with an adjustment for zero or negative values. The resulting transformed dataset is saved as Energydata.tx. This updated dataset encompasses both the normalized columns and the original columns that were either already normally distributed or couldn't

undergo testing due to insufficient data. The structure of Energydata.tx is then exhibited, presenting the initial rows of the processed dataset.

Ques 5. What are the correlations to the response variable (Total_Power) for energy.no.df? Which features exhibit a correlation above 0.6 to the response variable?

```
##           X1           Y1           X2           Y2           X3           Y3
##  0.413485848 -0.130783279  0.151053907  0.199538580 -0.001301448  0.430798248
##           X4           Y4           X5           Y5           X6           Y6
## -0.116242614  0.501398724  0.288452067 -0.167997751  0.196189700 -0.205474857
##           X7           Y7           X8           Y8           X9           Y9
##  0.125139024 -0.164219735  0.297495808 -0.197593522  0.312485451 -0.215477954
##          X10          Y10          X11          Y11          X12          Y12
##  0.300317817 -0.204493277  0.211829638 -0.185691985  0.145369899 -0.103416390
##          X13          Y13          X14          Y14          X15          Y15
##  0.076612258 -0.149390761 -0.004537084 -0.066707267  0.159709231 -0.126025178
##          X16          Y16          X17          Y17          X18          Y18
##  0.058974175 -0.044639983  0.279603309 -0.256553767  0.205994543 -0.154587890
##          X19          Y19          X20          Y20          X21          Y21
##  0.053121190 -0.060428292 -0.072902938 -0.044325599 -0.049224936 -0.018305947
##          X22          Y22          X23          Y23          X24          Y24
##  0.093929959 -0.050995511  0.031705332  0.010445017 -0.088333050  0.113074501
##          X25          Y25          X26          Y26          X27          Y27
##  0.062370266 -0.019366024 -0.038670690  0.071458571 -0.094550412  0.122894728
##          X28          Y28          X29          Y29          X30          Y30
## -0.152099434  0.193926889  0.076162826 -0.112327763 -0.003854492 -0.038259570
##          X31          Y31          X32          Y32          X33          Y33
## -0.147636951  0.050178264 -0.257974332  0.090642974  0.047650212 -0.126866191
##          X34          Y34          X35          Y35          X36          Y36
## -0.027937541 -0.142818209 -0.150630591 -0.067964058  0.064254978 -0.114833035
##          X37          Y37          X38          Y38          X39          Y39
##  0.054136310 -0.225978026  0.001688518 -0.177813530 -0.086554584 -0.151079155
##          X40          Y40          X41          Y41          X42          Y42
## -0.252518556  0.013708428 -0.181281781 -0.260546555 -0.190790318 -0.159685755
##          X43          Y43          X44          Y44          X45          Y45
## -0.103007111 -0.210744648 -0.133879328 -0.224709161  0.131556546 -0.530351242
##          X46          Y46          X47          Y47          X48          Y48
##  0.035514792 -0.558474628 -0.031670555 -0.386360061 -0.216891525 -0.374912072
##          X49          Y49       Power1       Power2       Power3       Power4
## -0.436232626 -0.197746157  0.437869925  0.434014887  0.133204364 -0.070058064
##       Power5       Power6       Power7       Power8       Power9      Power10
##  0.202793543  0.194966989  0.117305796  0.383119960  0.296417552  0.350384824
##      Power11      Power12      Power13      Power14      Power15      Power16
##  0.313931255  0.295167196  0.039895328  0.036690454  0.309328445  0.251308511
##      Power17      Power18      Power19      Power20      Power21      Power22
##  0.345842020  0.349424548  0.252722191  0.064947679  0.053252107  0.193596098
##      Power23      Power24      Power25      Power26      Power27      Power28
##  0.185688842  0.088826870  0.213533635  0.265848570  0.283430752  0.129707390
##      Power29      Power30      Power31      Power32      Power33      Power34
##  0.004285928  0.157746476  0.264640821  0.189578970  0.285849919  0.376404760
##      Power35      Power36      Power37      Power38      Power39      Power40
##  0.255984576  0.337559069  0.163686356  0.457198114  0.491762924  0.451697849
```

```
##      Power41      Power42      Power43      Power44      Power45      Power46
##  0.105580548  0.339434759  0.391268108  0.341634414  0.376892542  0.365059305
##      Power47      Power48      Power49           qW
##  0.402961754  0.205734901 -0.232437126  0.993712710
```

```
##        qW
## 0.9937127
```

The results display correlation coefficients between the variable "Total_Power" and all other variables in the Energydatanum.df dataset. Correlation coefficients, ranging from -1 to 1, signify the strength and direction of linear relationships—values closer to 1 or -1 indicate strong positive or negative correlations, while values near 0 indicate a weak linear relationship.

The Correleation_power object contains these coefficients, excluding the perfect correlation of "Total_Power" with itself. Variables with absolute correlation coefficients exceeding 0.6 are considered strongly correlated. Although the Corr_above_06 object doesn't currently show any output, if present, it would list the names of features most linearly associated with "Total_Power." These features could be crucial influencers or predictors in models involving "Total_Power" as a response variable. Notably, the high correlation value of 0.9504218960 with "qW" indicates a very strong linear relationship with "Total_Power."

Ques 6. Split each of the three data sets, energy.no.df, energy.df, and energy.tx into 70%/30% subsets, so you retain 30% for testing using random sampling without replacement. Call the data sets, energy.training and energy.testing, energy.no.training and energy.no.testing, and energy.tx.training and energy.tx.testing.

The datasets are partitioned into training and testing sets using a 70:30 split, ensuring reproducibility by setting a seed. The split_data function randomly assigns 70% of the data for training and reserves 30% for testing. This procedure is executed on three separate datasets, generating unique training and testing sets for each. These sets play a crucial role in the subsequent processes of model training and evaluation.

Ques 7. Build three Multiple Regression models. One for energy.training, energy.no.training, and energy.tx.training using backward elimination based on p-value for predicting Total_Power. Can you automate the elimination using loops?

```
##
## Call:
## lm(formula = formula, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2693.15  -105.49    14.02   131.04  2829.93
##
## Coefficients:
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept)  4.826e+03  1.993e+02   24.211  < 2e-16 ***
## X1          -3.973e-02  1.761e-02   -2.255 0.024119 *
## Y1          -4.762e-01  6.976e-02   -6.827 8.89e-12 ***
## X2           2.721e-01  1.948e-02   13.969  < 2e-16 ***
## Y2           5.821e-01  4.240e-02   13.728  < 2e-16 ***
```

```
## Y3           4.036e-01  7.083e-02    5.699 1.22e-08 ***
## X4          -2.070e-01  1.496e-02  -13.834  < 2e-16 ***
## X5          -5.247e-02  1.607e-02   -3.265 0.001096 **
## Y5          -1.276e+00  7.101e-02  -17.972  < 2e-16 ***
## X6           1.350e-01  1.799e-02    7.505 6.37e-14 ***
## Y6          -2.173e+00  8.904e-02  -24.402  < 2e-16 ***
## X7          -1.353e-01  1.656e-02   -8.169 3.24e-16 ***
## Y7           2.256e+00  9.770e-02   23.093  < 2e-16 ***
## X8           8.626e-02  1.534e-02    5.623 1.89e-08 ***
## Y8           5.817e-01  5.320e-02   10.935  < 2e-16 ***
## X9           7.745e-02  1.386e-02    5.590 2.30e-08 ***
## Y9           2.185e-01  4.948e-02    4.417 1.01e-05 ***
## X10          1.818e-01  1.454e-02   12.498  < 2e-16 ***
## X11          5.878e-02  1.604e-02    3.665 0.000248 ***
## Y11          2.119e-01  5.265e-02    4.024 5.74e-05 ***
## X12          3.639e-02  1.499e-02    2.427 0.015232 *
## Y12         -8.127e-01  5.124e-02  -15.859  < 2e-16 ***
## X13         -4.022e-02  1.441e-02   -2.790 0.005267 **
## Y13          3.779e-01  6.167e-02    6.128 9.02e-10 ***
## X14          7.801e-02  1.574e-02    4.958 7.18e-07 ***
## Y14          2.134e-01  6.429e-02    3.319 0.000903 ***
## X15          6.616e-02  1.441e-02    4.591 4.42e-06 ***
## Y15          1.809e-01  4.792e-02    3.775 0.000161 ***
## Y16         -5.821e-01  4.236e-02  -13.743  < 2e-16 ***
## Y17          9.022e-01  4.881e-02   18.483  < 2e-16 ***
## Y18         -2.445e-01  4.844e-02   -5.047 4.52e-07 ***
## Y19         -4.075e-01  5.117e-02   -7.963 1.75e-15 ***
## X20          3.478e-02  1.277e-02    2.724 0.006456 **
## Y20          1.948e-01  4.291e-02    4.539 5.68e-06 ***
## X21          4.135e-02  1.285e-02    3.219 0.001288 **
## Y21         -2.045e-01  4.086e-02   -5.004 5.64e-07 ***
## X22         -8.186e-02  1.340e-02   -6.109 1.02e-09 ***
## Y22          4.765e-01  5.661e-02    8.418  < 2e-16 ***
## X23         -5.029e-02  1.345e-02   -3.740 0.000184 ***
## Y23          1.430e-01  5.396e-02    2.651 0.008028 **
## X24          4.919e-02  1.328e-02    3.703 0.000213 ***
## Y24         -5.567e-01  5.043e-02  -11.038  < 2e-16 ***
## X25         -8.627e-02  1.325e-02   -6.513 7.49e-11 ***
## Y25          2.687e-01  4.184e-02    6.421 1.37e-10 ***
## X26         -7.120e-02  1.372e-02   -5.188 2.14e-07 ***
## X27         -7.916e-02  1.336e-02   -5.926 3.15e-09 ***
## Y27         -5.742e-01  5.090e-02  -11.281  < 2e-16 ***
## Y28          2.352e-01  5.094e-02    4.617 3.91e-06 ***
## X29         -4.214e-02  1.216e-02   -3.465 0.000531 ***
## X30         -6.988e-02  1.270e-02   -5.504 3.74e-08 ***
## X31          1.175e-01  1.370e-02    8.581  < 2e-16 ***
## Y31          3.090e-01  5.730e-02    5.392 7.03e-08 ***
## X32         -6.290e-02  1.336e-02   -4.708 2.52e-06 ***
## Y32         -6.534e-01  6.076e-02  -10.754  < 2e-16 ***
## X33          4.222e-02  1.316e-02    3.208 0.001336 **
## Y33         -4.014e-01  4.884e-02   -8.219  < 2e-16 ***
## X34          7.003e-02  1.369e-02    5.115 3.16e-07 ***
## Y34          4.925e-01  6.084e-02    8.095 6.00e-16 ***
## X35          6.010e-02  1.317e-02    4.564 5.04e-06 ***
```

```
## X36          1.055e-01  1.287e-02    8.196 2.61e-16 ***
## Y36          9.059e-01  5.783e-02   15.666  < 2e-16 ***
## X37         -7.916e-02  1.219e-02   -6.496 8.39e-11 ***
## Y37         -1.968e-01  4.989e-02   -3.945 8.01e-05 ***
## X38          7.096e-02  1.242e-02    5.712 1.13e-08 ***
## Y38          1.097e+00  5.283e-02   20.765  < 2e-16 ***
## X39         -1.553e-01  1.284e-02  -12.101  < 2e-16 ***
## Y39         -4.019e-01  5.443e-02   -7.382 1.60e-13 ***
## X40          1.128e-01  1.175e-02    9.595  < 2e-16 ***
## X41          5.874e-02  1.187e-02    4.947 7.58e-07 ***
## Y41         -3.897e-01  5.802e-02   -6.717 1.89e-11 ***
## X42          1.003e-01  1.255e-02    7.997 1.33e-15 ***
## Y42         -8.391e-01  4.998e-02  -16.787  < 2e-16 ***
## X43          5.450e-02  1.361e-02    4.004 6.26e-05 ***
## Y43         -2.063e+00  7.992e-02  -25.817  < 2e-16 ***
## X44         -2.942e-01  1.362e-02  -21.594  < 2e-16 ***
## Y44          3.053e+00  7.655e-02   39.884  < 2e-16 ***
## X45         -4.776e-02  1.202e-02   -3.973 7.13e-05 ***
## Y45          3.743e-01  7.378e-02    5.073 3.93e-07 ***
## X46         -1.156e-01  1.094e-02  -10.568  < 2e-16 ***
## Y46          5.791e-01  7.140e-02    8.111 5.25e-16 ***
## X47         -1.559e-01  1.052e-02  -14.819  < 2e-16 ***
## X48          2.560e-01  1.131e-02   22.626  < 2e-16 ***
## Y48         -1.730e+00  7.075e-02  -24.449  < 2e-16 ***
## X49         -4.997e-01  1.081e-02  -46.226  < 2e-16 ***
## Y49          1.416e+00  2.729e-02   51.894  < 2e-16 ***
## Power1       1.003e+00  4.240e-04 2365.838  < 2e-16 ***
## Power2       9.940e-01  4.200e-04 2366.620  < 2e-16 ***
## Power3       9.978e-01  2.767e-04 3605.491  < 2e-16 ***
## Power4       1.004e+00  3.319e-04 3023.695  < 2e-16 ***
## Power5       1.002e+00  3.393e-04 2953.198  < 2e-16 ***
## Power6       9.904e-01  3.684e-04 2688.329  < 2e-16 ***
## Power7       1.002e+00  3.365e-04 2977.850  < 2e-16 ***
## Power8       9.959e-01  3.143e-04 3168.909  < 2e-16 ***
## Power9       9.983e-01  3.109e-04 3210.571  < 2e-16 ***
## Power10      9.954e-01  3.442e-04 2891.890  < 2e-16 ***
## Power11      9.972e-01  3.664e-04 2721.956  < 2e-16 ***
## Power12      9.971e-01  3.085e-04 3231.920  < 2e-16 ***
## Power13      1.002e+00  3.068e-04 3264.846  < 2e-16 ***
## Power14      9.982e-01  3.418e-04 2920.268  < 2e-16 ***
## Power15      9.972e-01  3.259e-04 3059.785  < 2e-16 ***
## Power16      9.981e-01  2.464e-04 4051.071  < 2e-16 ***
## Power17      9.991e-01  2.828e-04 3532.810  < 2e-16 ***
## Power18      9.998e-01  3.119e-04 3206.024  < 2e-16 ***
## Power19      9.970e-01  2.840e-04 3511.026  < 2e-16 ***
## Power20      9.989e-01  2.977e-04 3355.409  < 2e-16 ***
## Power21      9.953e-01  3.050e-04 3263.337  < 2e-16 ***
## Power22      1.001e+00  3.241e-04 3089.411  < 2e-16 ***
## Power23      1.001e+00  3.287e-04 3046.556  < 2e-16 ***
## Power24      9.983e-01  3.060e-04 3262.162  < 2e-16 ***
## Power25      9.984e-01  3.332e-04 2996.135  < 2e-16 ***
## Power26      9.998e-01  3.665e-04 2728.292  < 2e-16 ***
## Power27      9.982e-01  3.502e-04 2850.375  < 2e-16 ***
## Power28      9.973e-01  2.884e-04 3458.633  < 2e-16 ***
```

```
## Power29      1.002e+00  3.228e-04 3103.348  < 2e-16 ***
## Power30      1.001e+00  3.426e-04 2923.123  < 2e-16 ***
## Power31      9.958e-01  3.663e-04 2718.477  < 2e-16 ***
## Power32      9.996e-01  3.325e-04 3006.118  < 2e-16 ***
## Power33      9.966e-01  3.352e-04 2973.588  < 2e-16 ***
## Power34      9.966e-01  3.660e-04 2723.248  < 2e-16 ***
## Power35      9.996e-01  3.455e-04 2893.477  < 2e-16 ***
## Power36      9.966e-01  3.201e-04 3113.375  < 2e-16 ***
## Power37      9.978e-01  2.996e-04 3330.178  < 2e-16 ***
## Power38      9.950e-01  3.441e-04 2891.922  < 2e-16 ***
## Power39      9.966e-01  3.572e-04 2790.091  < 2e-16 ***
## Power40      9.957e-01  3.270e-04 3045.244  < 2e-16 ***
## Power41      9.964e-01  3.260e-04 3056.884  < 2e-16 ***
## Power42      9.926e-01  3.796e-04 2615.169  < 2e-16 ***
## Power43      9.959e-01  3.851e-04 2586.197  < 2e-16 ***
## Power44      9.904e-01  3.833e-04 2583.842  < 2e-16 ***
## Power45      9.906e-01  4.940e-04 2005.477  < 2e-16 ***
## Power46      9.956e-01  5.377e-04 1851.559  < 2e-16 ***
## Power47      9.926e-01  6.279e-04 1580.654  < 2e-16 ***
## Power48      9.995e-01  8.355e-04 1196.250  < 2e-16 ***
## Power49      1.004e+00  6.209e-04 1616.658  < 2e-16 ***
## qW           3.931e+03  5.970e+02    6.585 4.64e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 264.7 on 25048 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 4.009e+07 on 134 and 25048 DF,  p-value: < 2.2e-16
```

The output indicates that the final model for EnergyData.training has an extremely high multiple R-squared value of 0.9999, suggesting that almost all of the variability in the response variable "Total_Power" can be explained by the model. The F-statistic is significantly large with a p-value less than 2.2e-16, indicating that the model is statistically significant at explaining the variation in the response variable compared to a model with no predictors.

Ques 8. Provide an analysis of all 3 models (using their respective testing data sets), including Adjusted R-Squared and RMSE. Which of these models is the best? Why?

The Multiple Regression Model trained on EnergyData.training exhibits an exceptionally high Adjusted R-squared value of 0.9999953, indicating its ability to explain almost all the variance in the response variable "Total_Power." However, the relatively elevated RMSE of 264.2661 suggests the possibility of substantial errors in predictions.

In the case of the Multiple Regression Model trained on EnergyDatanum.training, it achieves a flawless Adjusted R-squared value of 1, signifying a perfect fit to the data. Additionally, it boasts an exceptionally low RMSE of 0.01947136, indicating highly accurate predictions with minimal error.

For the Multiple Regression Model trained on Energy.tx.training, despite having a slightly lower Adjusted R-squared value of 0.9944193 compared to the other two models, it still demonstrates a robust fit to the data. Remarkably, it has the lowest RMSE at 0.0006193743, indicating extremely precise predictions.

In summary, all models perform well, but the one trained on EnergyDatanum.training appears to showcase the most superior predictive performance based on these testing sets.