PES University, Bangalore

UE20CS312 - Data Analytics

Worksheet 2b : Multiple Linear Regression

Course Anchor : Dr. Gowri Srinivasa

Prepared by : Nishanth M S - nishanthmsathish.23@gmail.com

Multiple Linear Regression

Multiple Linear Regression (MLR) is a statistical technique that uses several explanatory variables to predict the outcome of response variable.The goal of MLR is to model **a linear relationship** between explanatory(independent) variables and response(dependent) variables.

Data Dictionary

The data required for this worksheet can be downloaded from this GitHub Link. The data was obtained from this dataset from Kaggle. The dataset contains features of songs on Spotify collected using Spotify API.The features are as follows :

-acousticness : A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic.

-danceability : Danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable.

-duration_ms : The duration of track in milliseconds.

-energy : Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity.Perceptual features contributing to this attribute include dynamic range, perceived loudness, timbre, onset rate, and general entropy.

-instrumentalness : Predicts whether a track contains no vocals.The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content. Values above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.0.

-key : The key the track is in. Integers map to pitches using standard Pitch Class notation

-liveness : Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides strong likelihood that the track is live.

-loudness : The overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track and are useful for comparing relative loudness of tracks. Loudness is the quality of a sound that is the primary psychological correlate of physical strength (amplitude). Values typical range between -60 and 0 db.

-mode : Mode indicates the modality (major or minor) of a track, the type of scale from which its melodic content is derived. Major is represented by 1 and minor is 0.

-speechiness : Speechiness detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and other non-speech-like tracks.

-tempo : The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration.

-time_signature : An estimated overall time signature of a track. The time signature (meter) is a notational convention to specify how many beats are in each bar (or measure).

-valence : A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry).

Throughout the course of this worksheet , our response variable is energy. We shall try and apply the concepts learnt in class to predict the energy of a song using the other features of a song.

Libraries used

-tidyverse

-corrplot

-olsrr : documentation

Points

The problems for this worksheet is for a total of 10 points and the weightage is not uniformly distributed.

- *Problem 1* : 0.5 points
- *Problem 2* : 2 points
- *Problem 3* : 2 points
- *Problem 4* : 1 point
- *Problem 5* : 1.5 points
- *Problem 6* : 1 point
- *Problem 7* : 2 points

Loading packages:

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.8     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```r
library(olsrr)
```

```
##
## Attaching package: 'olsrr'
##
## The following object is masked from 'package:datasets':
##
##     rivers
```

```r
library(ggpubr)
library(dplyr)
library(ggplot2)
library(broom)
library(car)
```

```
## Loading required package: carData
##
## Attaching package: 'car'
##
## The following object is masked from 'package:dplyr':
##
##     recode
##
## The following object is masked from 'package:purrr':
##
##     some
```

```r
library(AICcmodavg)
```

Loading the Dataset

After downloading the dataset and ensuring the working directory is right , we read the csv into the dataframe.

```r
library(tidyverse)
spotify_df <- read_csv("spotify.csv")
head(spotify_df)
```

```
## # A tibble: 6 x 13
##   danceabil~1 energy   key loudn~2  mode speec~3 acous~4 instr~5 liven~6 valence
##         <dbl>  <dbl> <dbl>   <dbl> <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1       0.803 0.624      7   -6.76     0  0.0477   0.451 7.34e-4  0.1     0.628
## 2       0.762 0.703     10   -7.95     0  0.306    0.206 0        0.0912  0.519
## 3       0.261 0.0149     1  -27.5      1  0.0419   0.992 8.97e-1  0.102   0.0382
## 4       0.722 0.736      3   -6.99     0  0.0585   0.431 1.18e-6  0.123   0.582
```

```
## 5       0.787 0.572      1   -7.52     1  0.222   0.145  0        0.0753  0.647
## 6       0.778 0.632      8   -6.42     1  0.125   0.0404 0        0.0912  0.827
## # ... with 3 more variables: tempo <dbl>, duration_ms <dbl>,
## #   time_signature <dbl>, and abbreviated variable names 1: danceability,
## #   2: loudness, 3: speechiness, 4: acousticness, 5: instrumentalness,
## #   6: liveness
```

Problem-1 (0.5 Points)

Check for missing values in the dataset and normalize the dataset.

```
#Check for missing values in each column
sapply(spotify_df,anyNA)
```

```
##     danceability           energy              key         loudness
##            FALSE            FALSE            FALSE            FALSE
##             mode      speechiness     acousticness instrumentalness
##            FALSE            FALSE            FALSE            FALSE
##         liveness          valence            tempo      duration_ms
##            FALSE            FALSE            FALSE            FALSE
##   time_signature
##            FALSE
```

```
#count of missing values in each column
sapply(spotify_df, function(x) sum(is.na(x)))
```

```
##     danceability           energy              key         loudness
##                0                0                0                0
##             mode      speechiness     acousticness instrumentalness
##                0                0                0                0
##         liveness          valence            tempo      duration_ms
##                0                0                0                0
##   time_signature
##                0
```

Problem-2 (2 Points)

Fit a linear model to predict the *energy* rating using *all* other attributes.Get the summary of the model and explain the results in detail.[*Hint* : Use the lm() function. Click here To get the documentation of the same.]

```
full_model<-lm(energy~.,data=spotify_df)
summary(full_model)
```

```
##
## Call:
## lm(formula = energy ~ ., data = spotify_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.26070 -0.05953 -0.00253  0.07230  0.32407
```

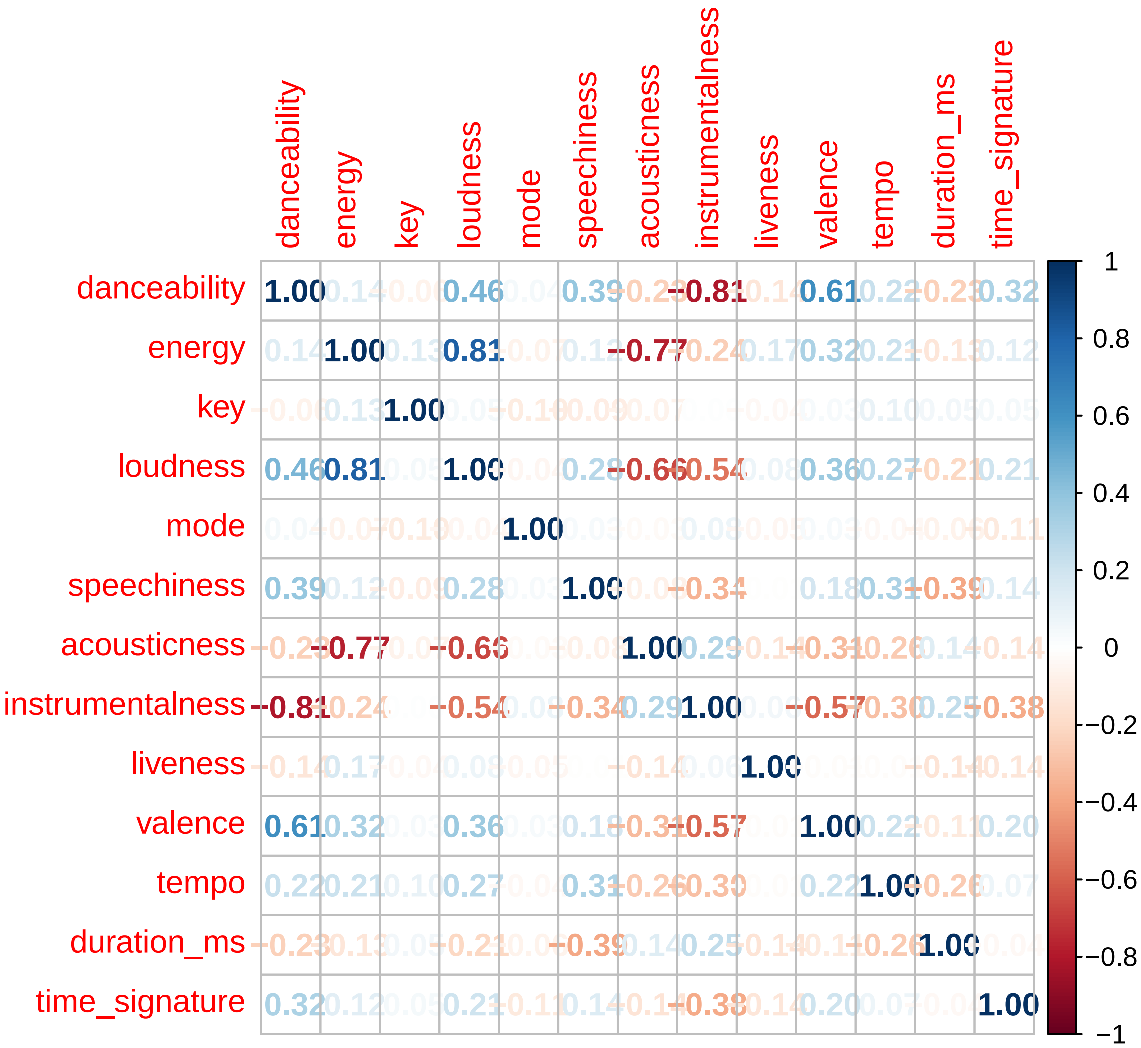
```
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.048e+00  1.070e-01   9.787  < 2e-16 ***
## danceability     -3.303e-01  6.670e-02  -4.952 1.67e-06 ***
## key               3.785e-03  2.291e-03   1.652  0.10030
## loudness          2.796e-02  1.818e-03  15.381  < 2e-16 ***
## mode             -2.495e-02  1.579e-02  -1.580  0.11582
## speechiness       5.095e-02  7.600e-02   0.670  0.50343
## acousticness     -2.785e-01  3.354e-02  -8.306 2.21e-14 ***
## instrumentalness  1.121e-01  4.189e-02   2.677  0.00811 **
## liveness          4.919e-02  7.610e-02   0.646  0.51880
## valence           1.988e-01  3.774e-02   5.269 3.85e-07 ***
## tempo            -2.218e-04  3.052e-04  -0.727  0.46817
## duration_ms      -6.723e-08  1.191e-07  -0.565  0.57298
## time_signature    1.388e-02  1.856e-02   0.748  0.45535
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.106 on 182 degrees of freedom
## Multiple R-squared:  0.844,  Adjusted R-squared:  0.8338
## F-statistic: 82.08 on 12 and 182 DF,  p-value: < 2.2e-16
```

Summary:The attributes danceability,loudness,acousticness,instrumentalness and valence are all significant. That is they are important predictors in determining energy with alpha set to 0.05. Larger the adjusted R-squared and smaller the residual standards error implies better the model. R-squared value is always less than adjusted R-squared value. F-statistic is analysis of whole model,in this case not all beta are zero.

Problem-3 (2 points)

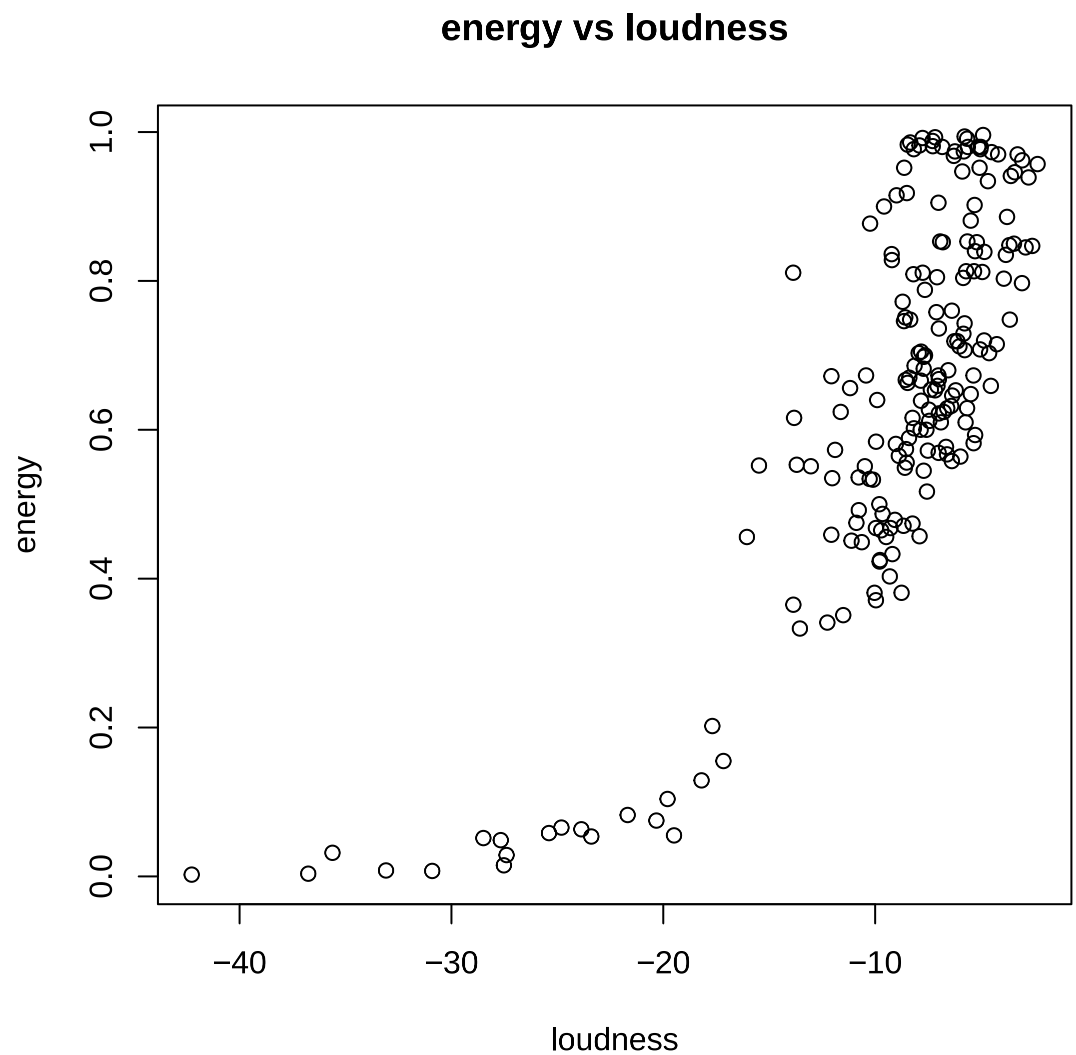
With the help of a correlogram and scatter plots, choose the features you think are important and model an MLR. Justify your choice and explain the new findings.

```
correlation<-cor(spotify_df)
corrplot(correlation,method = 'number')
```



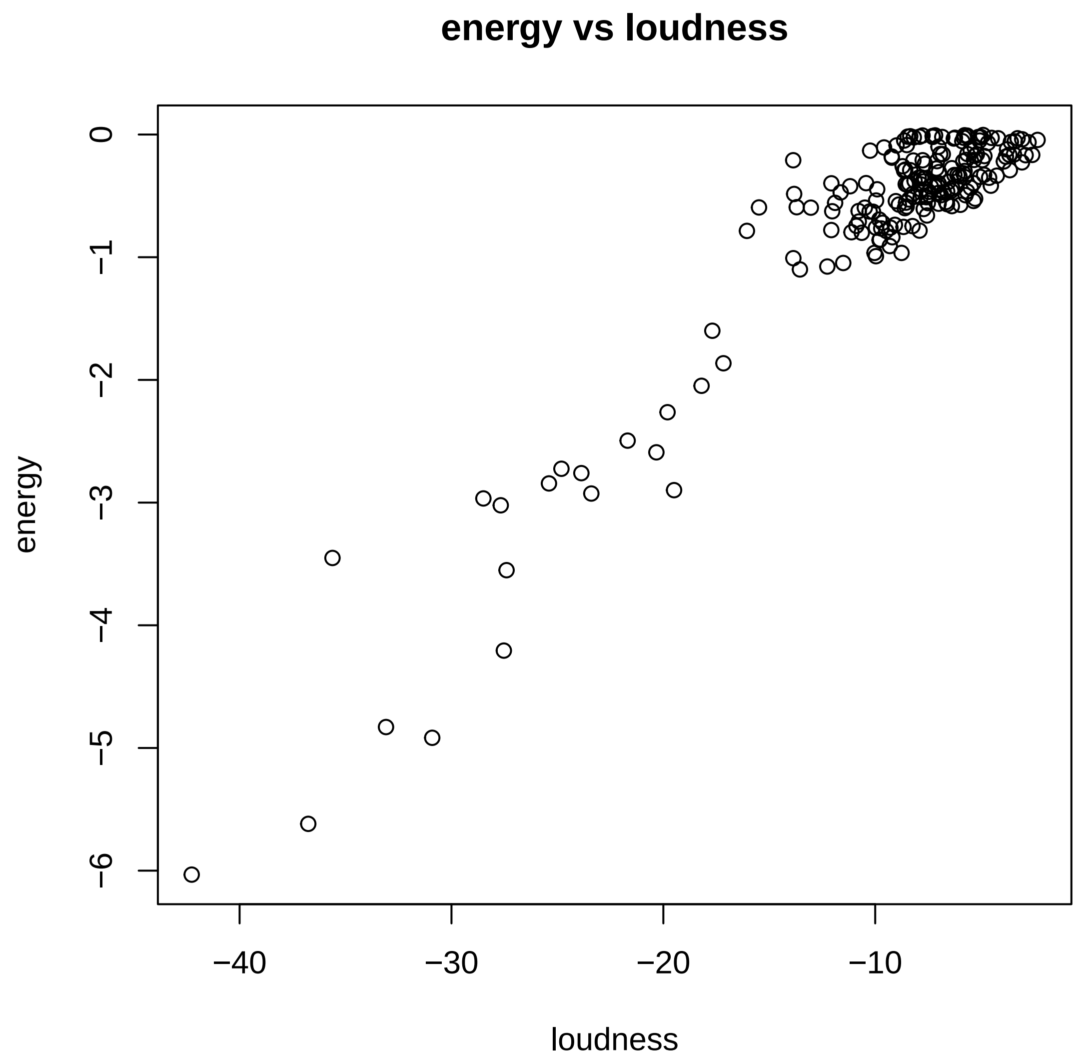
Energy is strongly positively correlated to loudness and accoustiness. And accoustiness and loudness are also negatively correlated.Implies comparing energy and loudness is sufficient to draw insights about accoustiness.

```
plot(x=spotify_df$loudness,y=spotify_df$energy,xlab="loudness",ylab="energy",main="energy vs loudness")
```

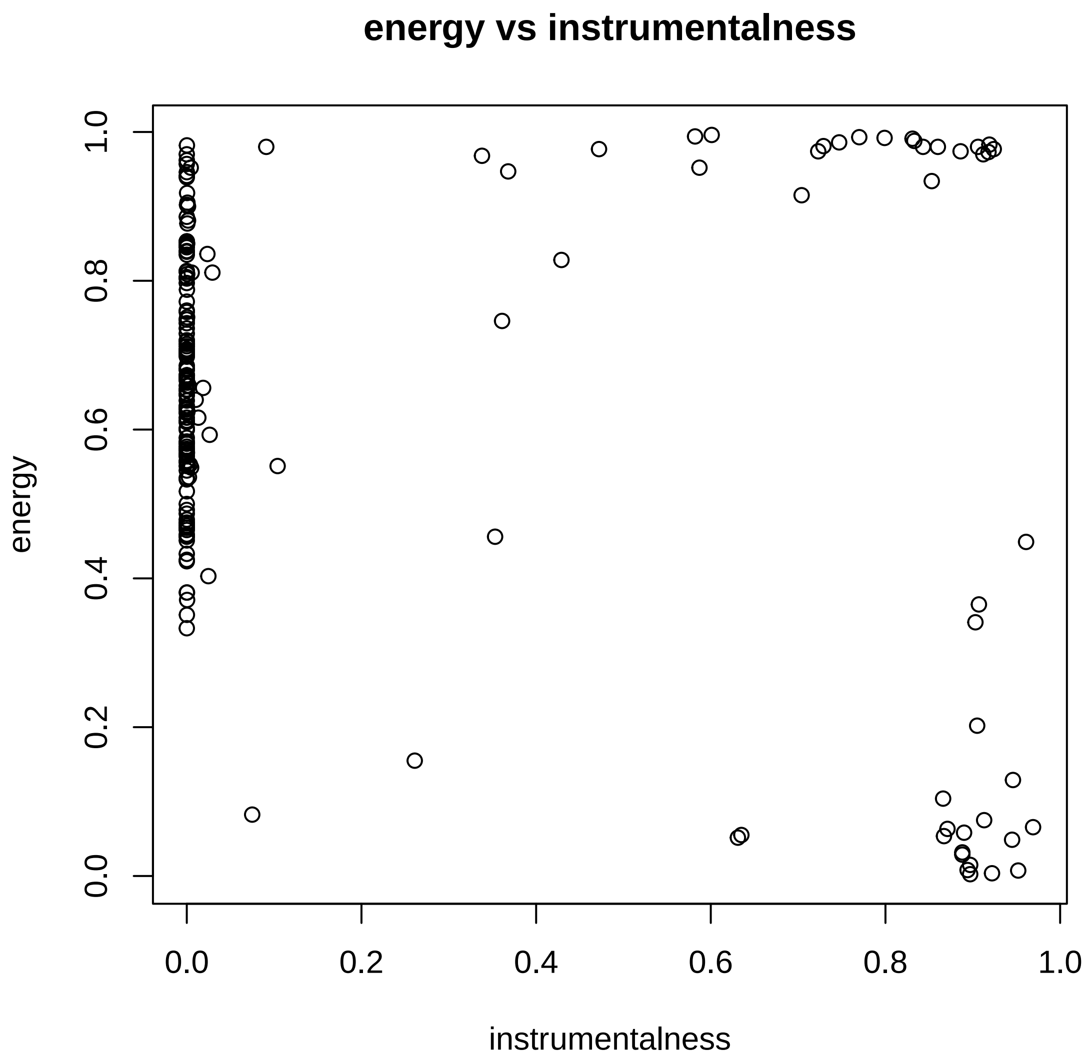


ln(energy) vs x

```
plot(x=spotify_df$loudness,y=log(spotify_df$energy),xlab="loudness",ylab="energy",main="energy vs loudne
```

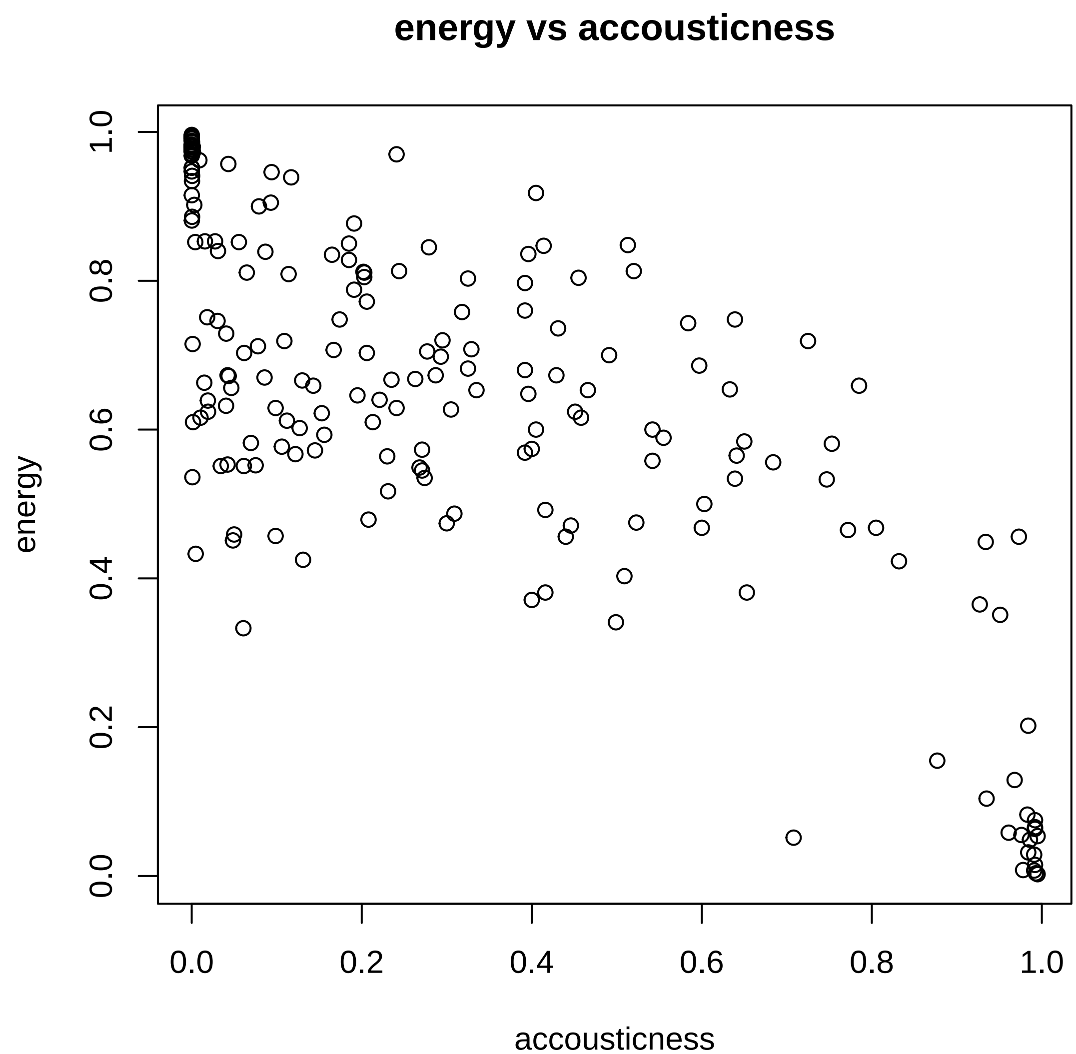



```
plot(x=spotify_df$instrumentalness,y=spotify_df$energy,xlab="instrumentalness",ylab="energy",main="energ
```



bad
model

```
plot(x=spotify_df$acousticness,y=spotify_df$energy,xlab="accousticness",ylab="energy",main="energy vs ac
```




```
reduced_model<-lm(energy~loudness+acousticness,data=spotify_df)
summary(reduced_model)
```

```
##
## Call:
## lm(formula = energy ~ loudness + acousticness, data = spotify_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.31751 -0.08934  0.00034  0.09070  0.29379
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.948987   0.016278  58.300  < 2e-16 ***
## loudness      0.021425   0.001895  11.308  < 2e-16 ***
## acousticness -0.336616   0.038541  -8.734  1.2e-15 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1286 on 192 degrees of freedom
## Multiple R-squared:  0.758,  Adjusted R-squared:  0.7555
## F-statistic: 300.7 on 2 and 192 DF,  p-value: < 2.2e-16
```

Since adjusted Rsquared value is higher implies its a good model.

Problem-4 (1 Point)

Conduct a partial F-test to determine if the attributes not chosen by you in *Problem-3* are significant to predict the energy.What are the null and alternate hypotheses? [*Hint* : Use the anova function between models created in *Problem-2* and *Problem-3*]

```
anova(reduced_model,full_model)
```

```
## Analysis of Variance Table
##
## Model 1: energy ~ loudness + acousticness
## Model 2: energy ~ danceability + key + loudness + mode + speechiness +
##     acousticness + instrumentalness + liveness + valence + tempo +
##     duration_ms + time_signature
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    192 3.1756
## 2    182 2.0469 10    1.1288 10.037 2.416e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Rejecting null hypothesis(since p value <0.05).Among the parameters left out,they are significant ## Problem-5 (1.5 Points)

AIC - Akaike Information Criterion is used to compare different models and determine the best fit for the data. The best-fit model according to AIC is the one that explains greatest amount of variation using the fewest number of attributes. Check this resource to learn more about AIC.

Build a model based on AIC using Stepwise AIC regression.Elucidate your observations from the new model. (*Hint* : Use an appropriate function in olsrr package.)

```
full_model<-lm(energy~.,data=spotify_df)
stepwise_model<-lm(energy~loudness+acousticness+danceability+valence+instrumentalness+mode+key,data=spo
reduced_model<-lm(energy~loudness+acousticness,data=spotify_df)

models<- list(full_model,reduced_model,stepwise_model)
models.names<- c('full_model','reduced_model','stepwise_model')
aictab(cand.set=models,modnames=models.names)
```

```
##
## Model selection based on AICc:
##
##                 K    AICc Delta_AICc AICcWt Cum.Wt     LL
## stepwise_model  9 -313.64       0.00   0.99   0.99 166.30
## full_model     14 -304.84       8.80   0.01   1.00 167.58
## reduced_model   4 -241.31      72.32   0.00   1.00 124.76
```

Analysis:Lower the value of AIC, better is the model.Therfore stepwise model is the best amongst full model and reduced model.

Problem-6 (1 Point)

Plot the residuals of the models built till now and comment on it satisfying the assumptions of MLR.

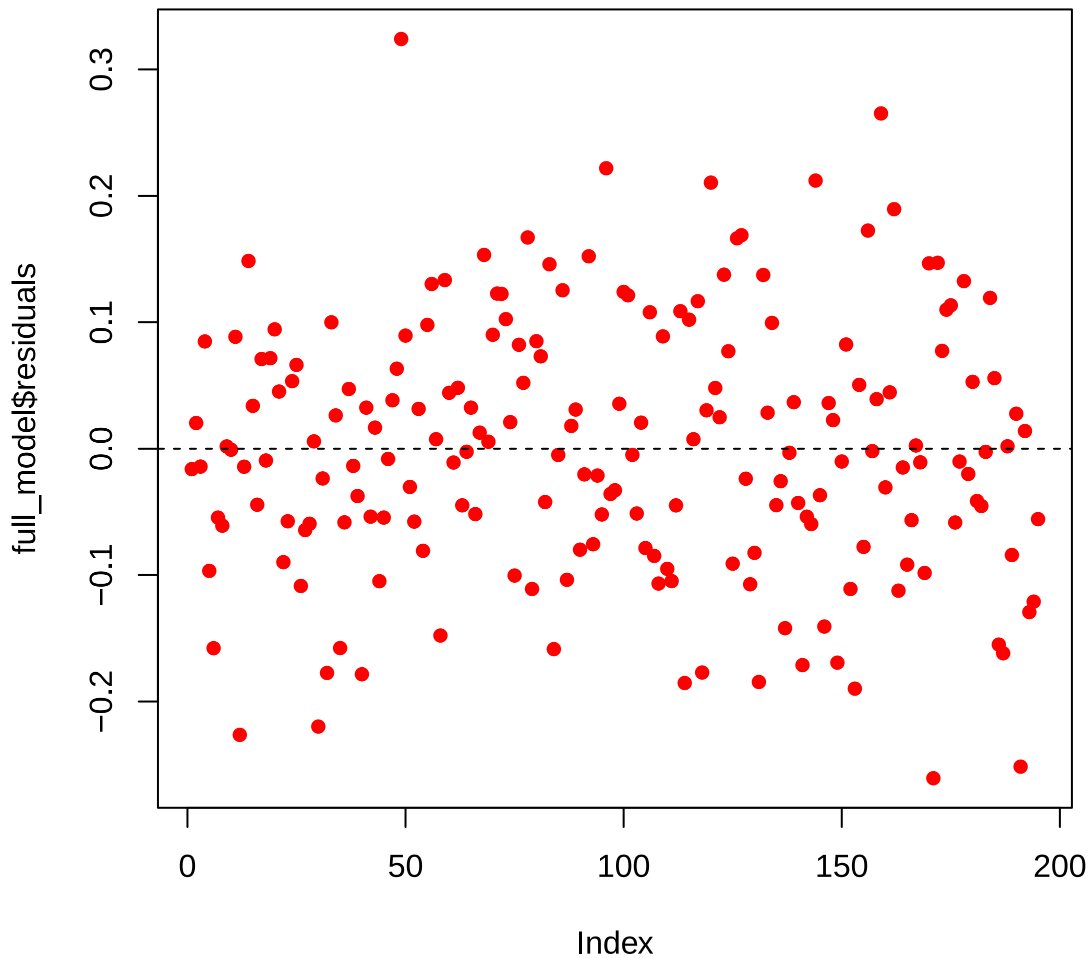
```
stepwise_model<-lm(energy~loudness+acousticness+danceability+valence+instrumentalness+mode+key,data=spo
summary(stepwise_model)
```

```
##
## Call:
## lm(formula = energy ~ loudness + acousticness + danceability +
##     valence + instrumentalness + mode + key, data = spotify_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.27482 -0.06470 -0.00293  0.07264  0.32765
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.077529   0.049529  21.755  < 2e-16 ***
## loudness          0.028203   0.001779  15.856  < 2e-16 ***
## acousticness     -0.277290   0.032473  -8.539 4.63e-15 ***
## danceability     -0.321907   0.063730  -5.051 1.04e-06 ***
## valence           0.194643   0.037161   5.238 4.35e-07 ***
## instrumentalness  0.106548   0.040199   2.650  0.00873 **
## mode             -0.025309   0.015532  -1.629  0.10491
## key               0.003418   0.002247   1.521  0.12988
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1053 on 187 degrees of freedom
## Multiple R-squared:  0.842,  Adjusted R-squared:  0.8361
## F-statistic: 142.3 on 7 and 187 DF,  p-value: < 2.2e-16
```

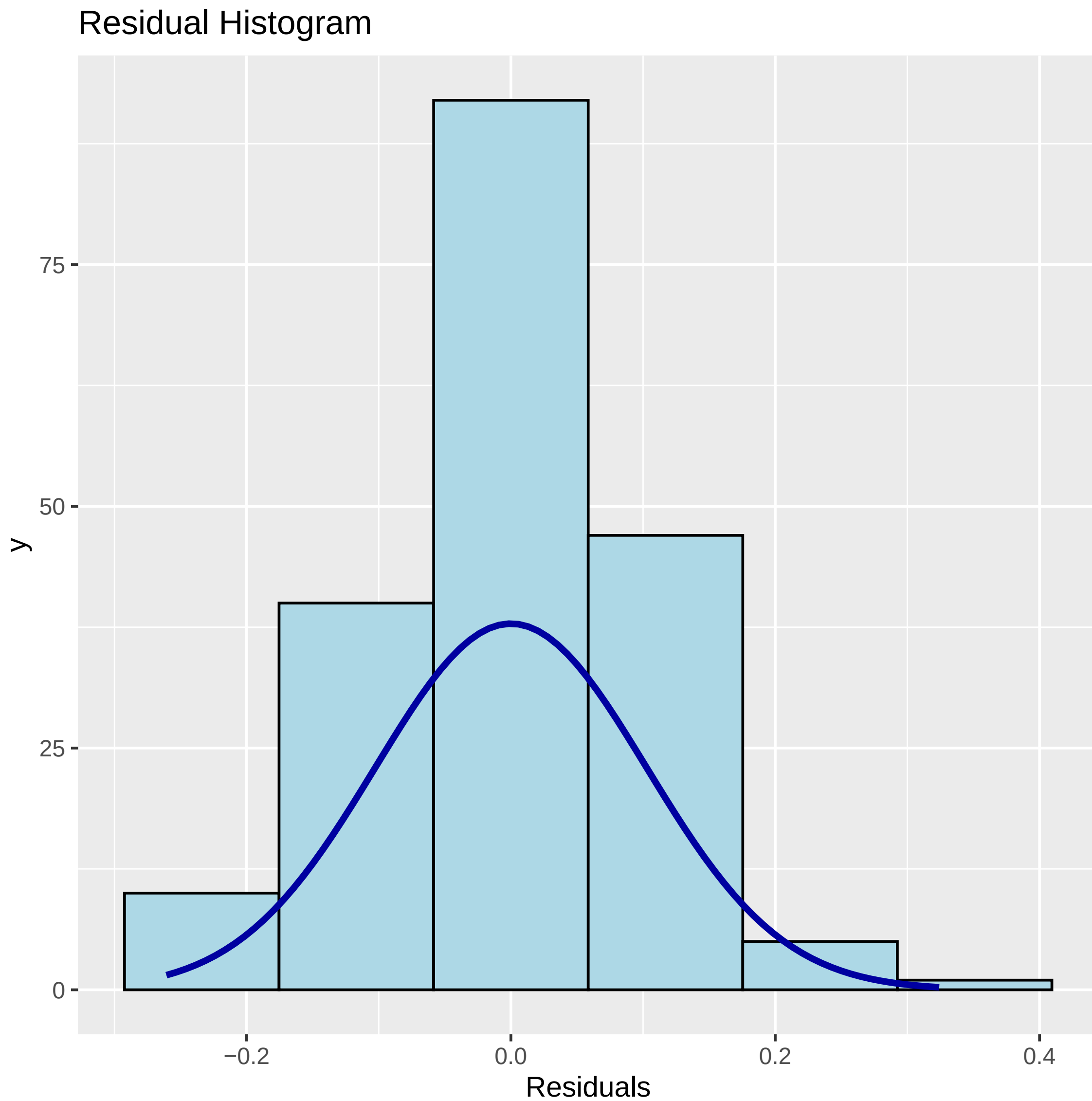
```
print("full model residuals")
```

```
## [1] "full model residuals"
```

```
plot(full_model$residuals,pch=16,col="red")
abline(h=0,lty=2)
```




```
ols_plot_resid_hist(full_model)
```

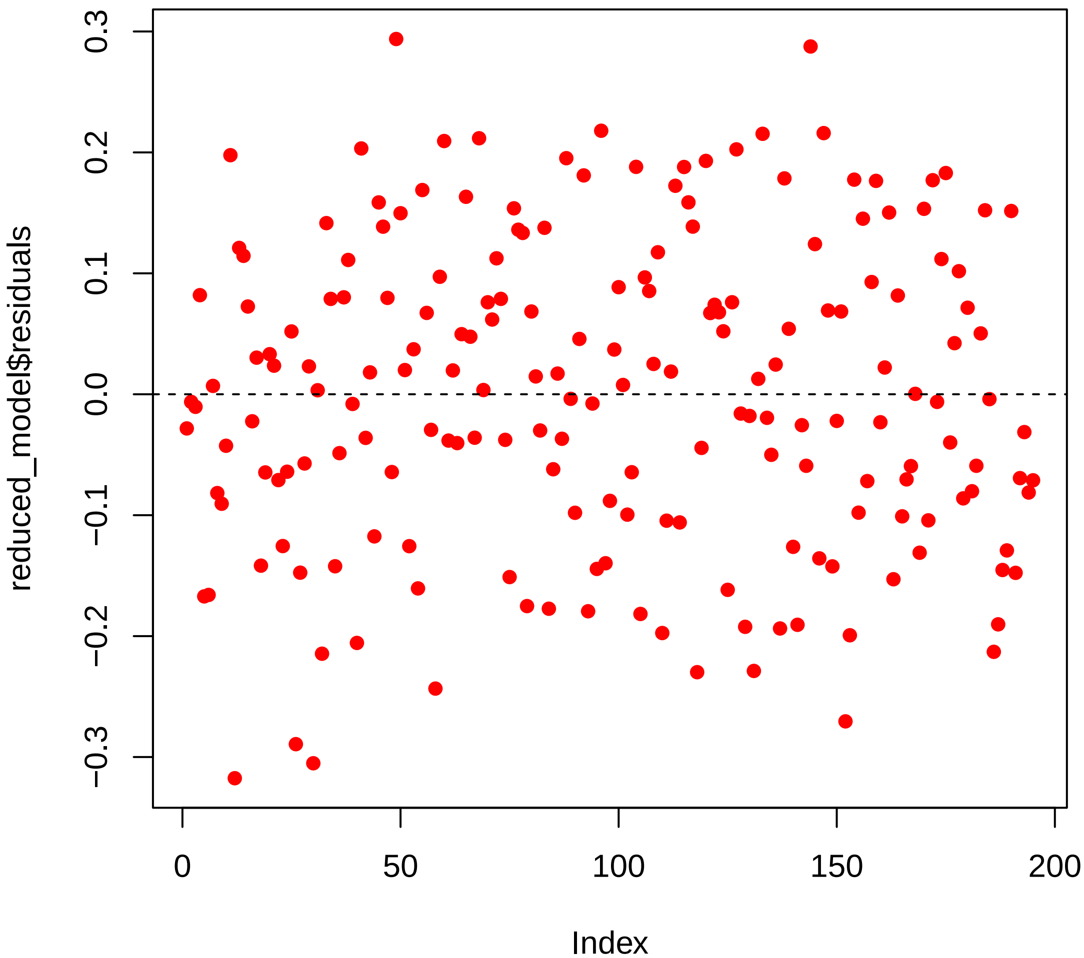


Histogram appears to be normally distributed.

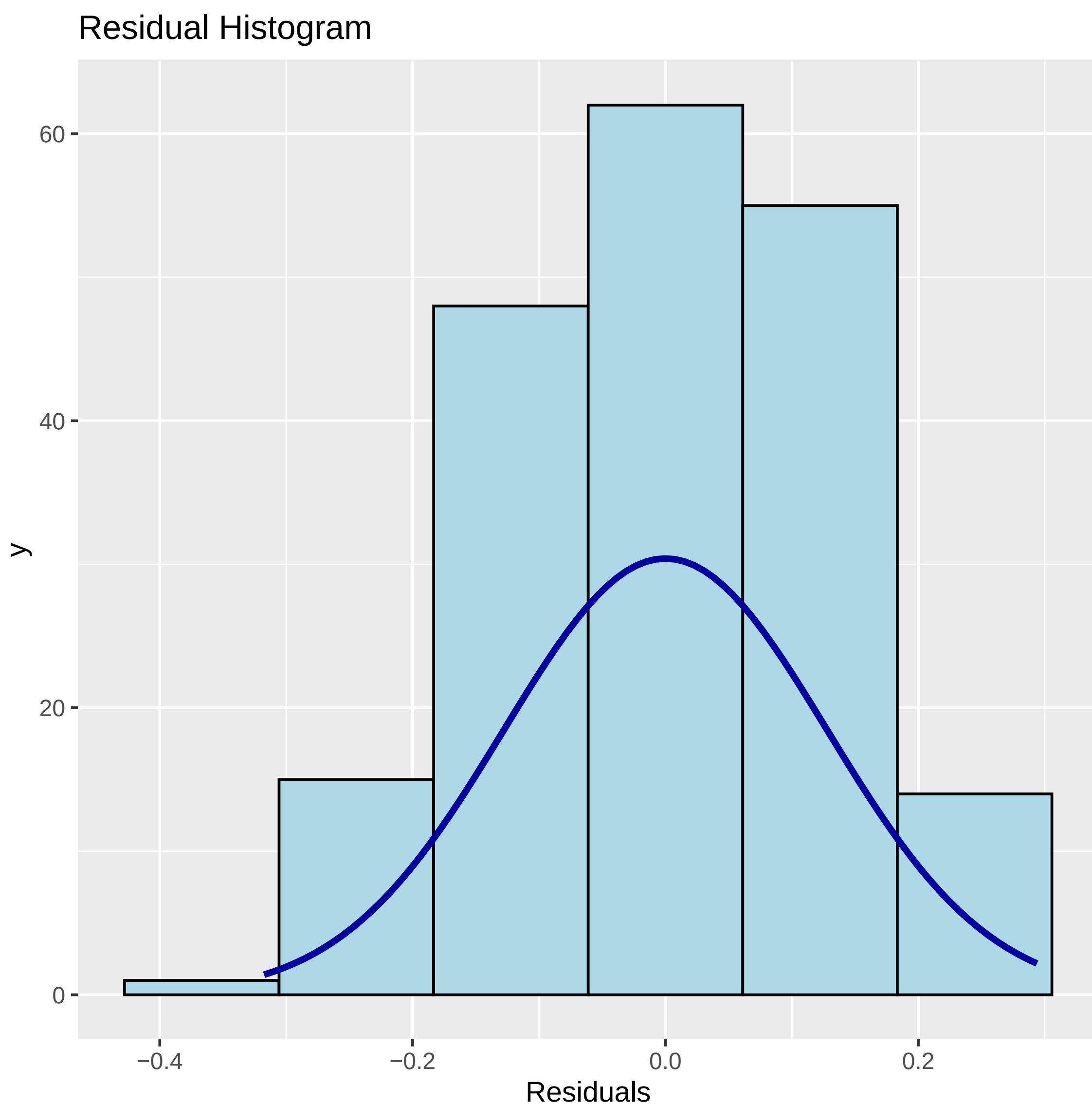
```
print("reduced residuals plots")
```

```
## [1] "reduced residuals plots"
```

```
plot(reduced_model$residuals,pch=16,col="red")
abline(h=0,lty=2)
```



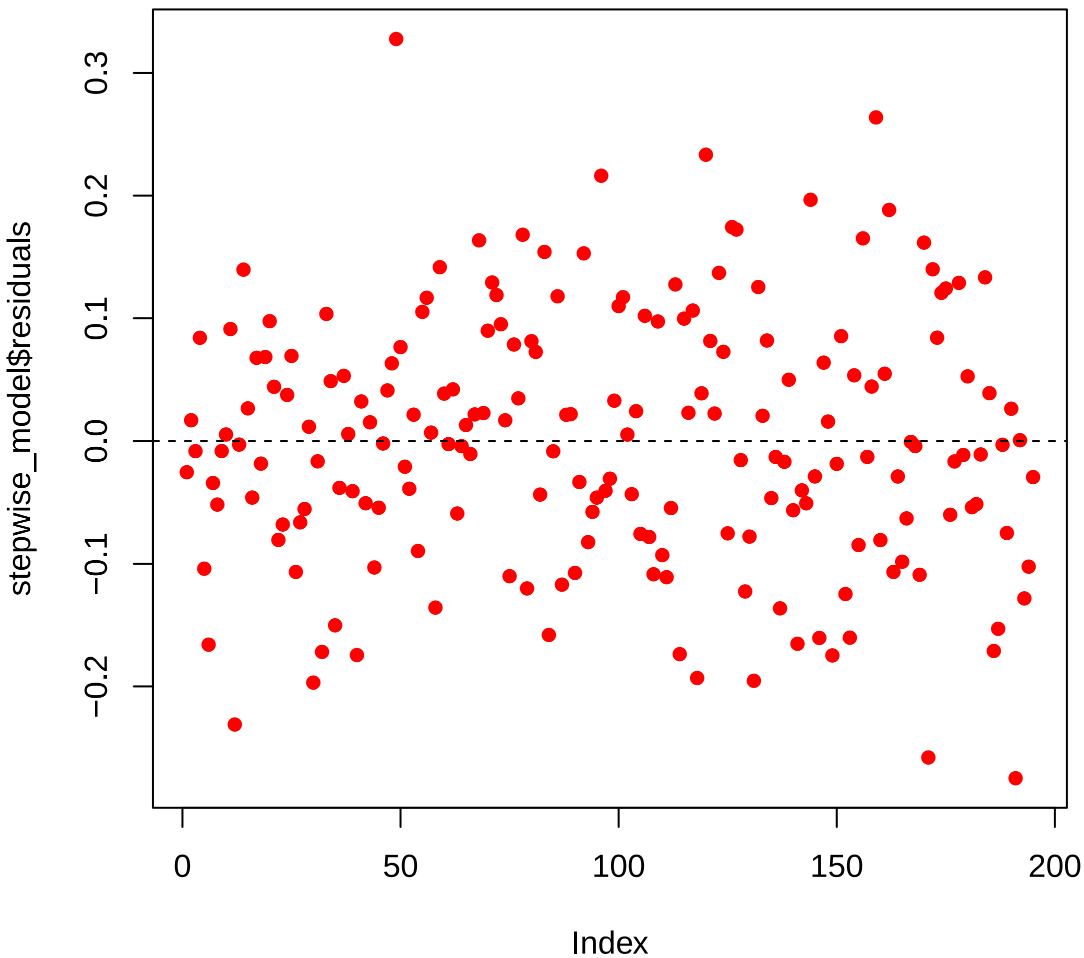

```
ols_plot_resid_hist(reduced_model)
```



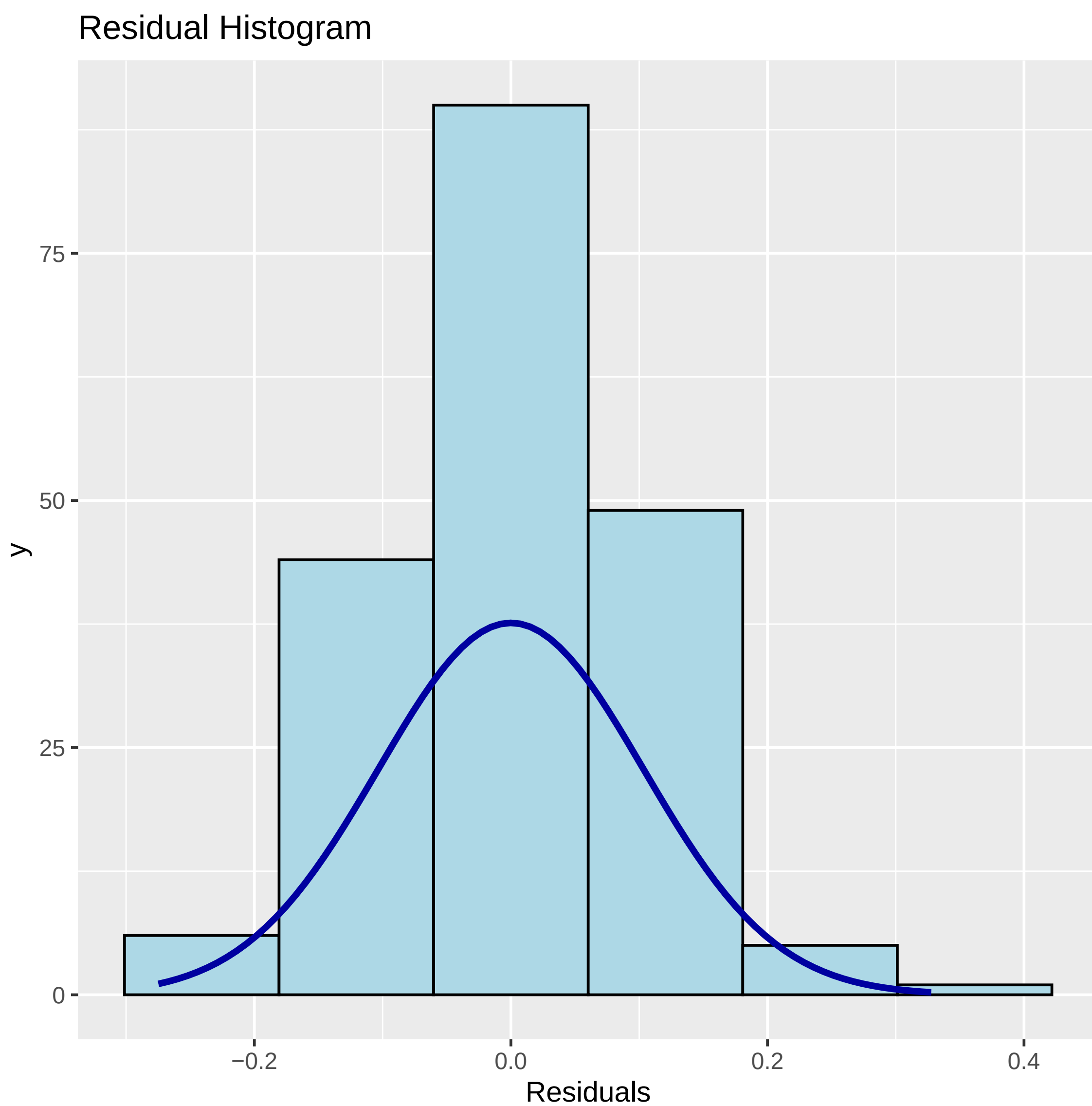

```
print("stepwise residuals plots")
```

```
## [1] "stepwise residuals plots"
```

```
plot(stepwise_model$residuals,pch=16,col="red")
abline(h=0,lty=2)
```




```
ols_plot_resid_hist(stepwise_model)
```



Comment:Stepwise model is better compared to full model and reduced model.

Problem-7 (2 Points)

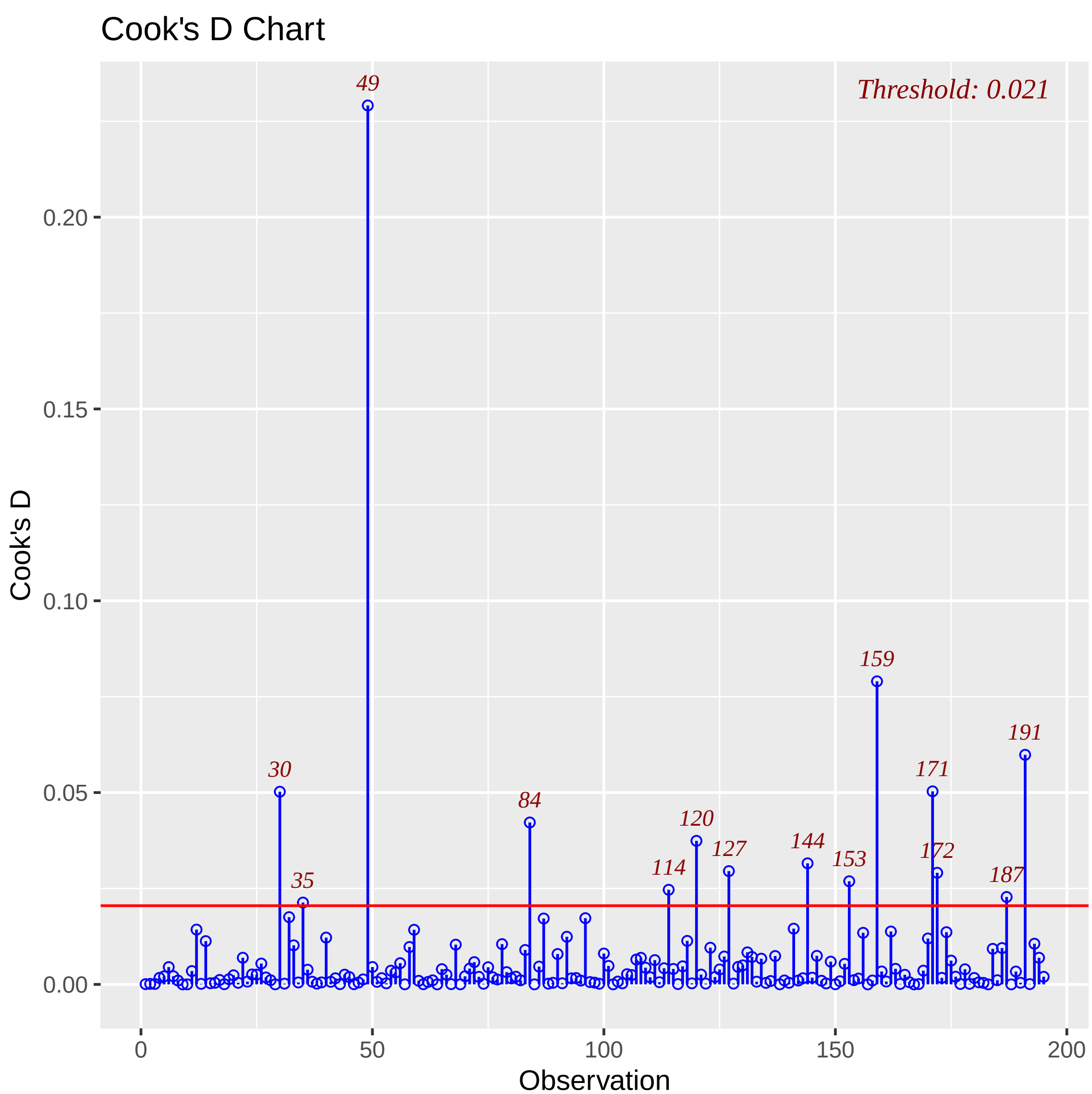
For the model built in ***Problem-2*** , determine the presence of multicollinearity using VIF. Determine if there are outliers in the data using Cook's Distance. If you find any , remove the outliers and fit the model for *Problem-2* and see if the fit improves. [*Hint* : All the relevant functions can be found in *olsrr* package. An observation can be termed as an outlier if it has a Cook's distance of more than 4/n where n is the number of records.]

```
ols_vif_tol(full_model)
```

```
##          Variables Tolerance      VIF
## 1     danceability 0.2776703 3.601393
## 2              key 0.9467671 1.056226
## 3         loudness 0.4119898 2.427245
```

```
## 4             mode 0.9308390 1.074300
## 5       speechiness 0.6921660 1.444740
## 6      acousticness 0.5009458 1.996224
## 7  instrumentalness 0.2755568 3.629016
## 8          liveness 0.8914397 1.121781
## 9           valence 0.5680642 1.760364
## 10            tempo 0.7892957 1.266952
## 11      duration_ms 0.7855373 1.273014
## 12   time_signature 0.8262918 1.210226
```

```
cookd<-ols_plot_cooksd_chart(full_model)
```



Nothing is violating threshold(value of threshold being 4/n).

#remove outliers

```r
new_df<-spotify_df[-c(30,35,79,84,114,120,127,144,153,159,171,172,187,191),]
new_full_model<-lm(energy~.,data=new_df)
summary(new_full_model)
```

```
##
## Call:
## lm(formula = energy ~ ., data = new_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.20976 -0.05895  0.00150  0.06087  0.37608
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.015e+00  1.002e-01  10.126  < 2e-16 ***
## danceability     -3.066e-01  6.692e-02  -4.581 8.96e-06 ***
## key               5.700e-03  2.103e-03   2.711 0.007413 **
## loudness          2.993e-02  1.766e-03  16.947  < 2e-16 ***
## mode             -1.632e-02  1.464e-02  -1.115 0.266502
## speechiness       2.086e-02  7.368e-02   0.283 0.777410
## acousticness     -2.498e-01  3.176e-02  -7.865 4.27e-13 ***
## instrumentalness  1.454e-01  4.110e-02   3.537 0.000523 ***
## liveness          6.993e-02  7.112e-02   0.983 0.326934
## valence           1.852e-01  3.454e-02   5.362 2.69e-07 ***
## tempo            -2.027e-04  2.924e-04  -0.693 0.489226
## duration_ms      -1.580e-07  1.093e-07  -1.446 0.150056
## time_signature    2.272e-02  1.743e-02   1.303 0.194210
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09326 on 168 degrees of freedom
## Multiple R-squared:  0.8663, Adjusted R-squared:  0.8567
## F-statistic: 90.71 on 12 and 168 DF,  p-value: < 2.2e-16
```