UE20CS312 - Data Analytics

Worksheet 2b :Multiple Linear Regression

PES University

'SUNDEEP A, Dept. of CSE - PES1UG20CS445'

2022-09-14

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 the Dataset

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

```
library(tidyverse)
spotify_df <- read_csv('spotify.csv')
```

Problem-1 (0.5 Points)

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

```
colSums((is.na(spotify_df)))#displays the number of missing values in each column.
```

```
##     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
```

```r
#Since there are No Missing Values, We move on to Normalizing the dataset.

#Normalizing using Min-Max Scaling to Normalize all the data from 0 to 1
library(caret)
```

```
## Loading required package: lattice

##
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':
##
##     lift
```

```r
normalized <- preProcess(as.data.frame(spotify_df), method=c("range"))
spotify_df <- predict(normalized, as.data.frame(spotify_df))
head(spotify_df)
```

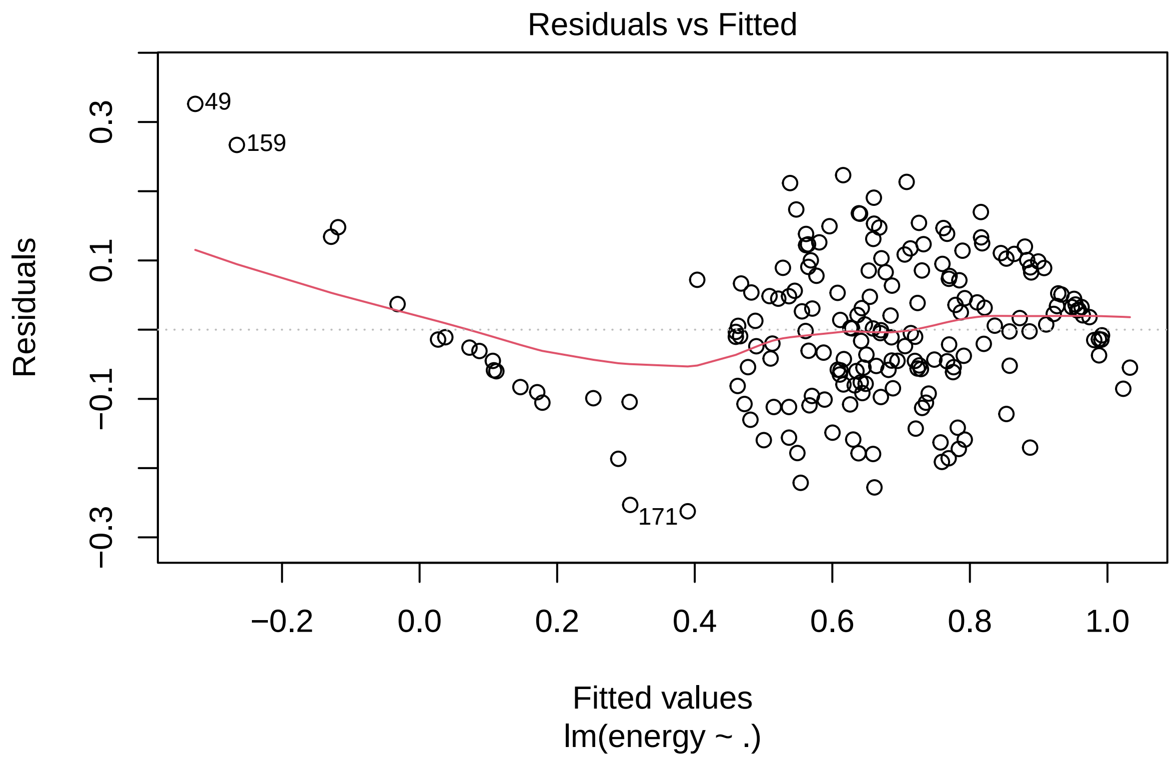
```
##   danceability     energy        key  loudness mode speechiness acousticness
## 1    0.8247549 0.62560386 0.63636364 0.8890920    0  0.03885201   0.45326466
## 2    0.7745098 0.70511272 0.90909091 0.8593613    0  0.54314721   0.20703275
## 3    0.1605392 0.01258052 0.09090909 0.3690169    1  0.02752831   0.99698492
## 4    0.7254902 0.73832528 0.27272727 0.8833312    0  0.05993752   0.43316409
## 5    0.8051471 0.57326892 0.09090909 0.8702567    1  0.37914877   0.14572602
## 6    0.7941176 0.63365539 0.72727273 0.8978334    1  0.18976962   0.04060007
##   instrumentalness   liveness     valence     tempo duration_ms time_signature
## 1     7.574819e-04 0.11151859 0.627394940 0.2986443   0.3932821           0.75
## 2     0.000000e+00 0.09684947 0.512014396 0.7605056   0.2940693           0.75
## 3     9.256966e-01 0.11485248 0.003069758 0.1261836   0.3629418           0.75
## 4     1.217750e-06 0.14985831 0.578702234 0.2476870   0.2278801           0.75
## 5     0.000000e+00 0.07034506 0.647507145 0.7921078   0.1768309           0.75
## 6     0.000000e+00 0.09684947 0.838043823 0.6739248   0.2540198           0.75
```

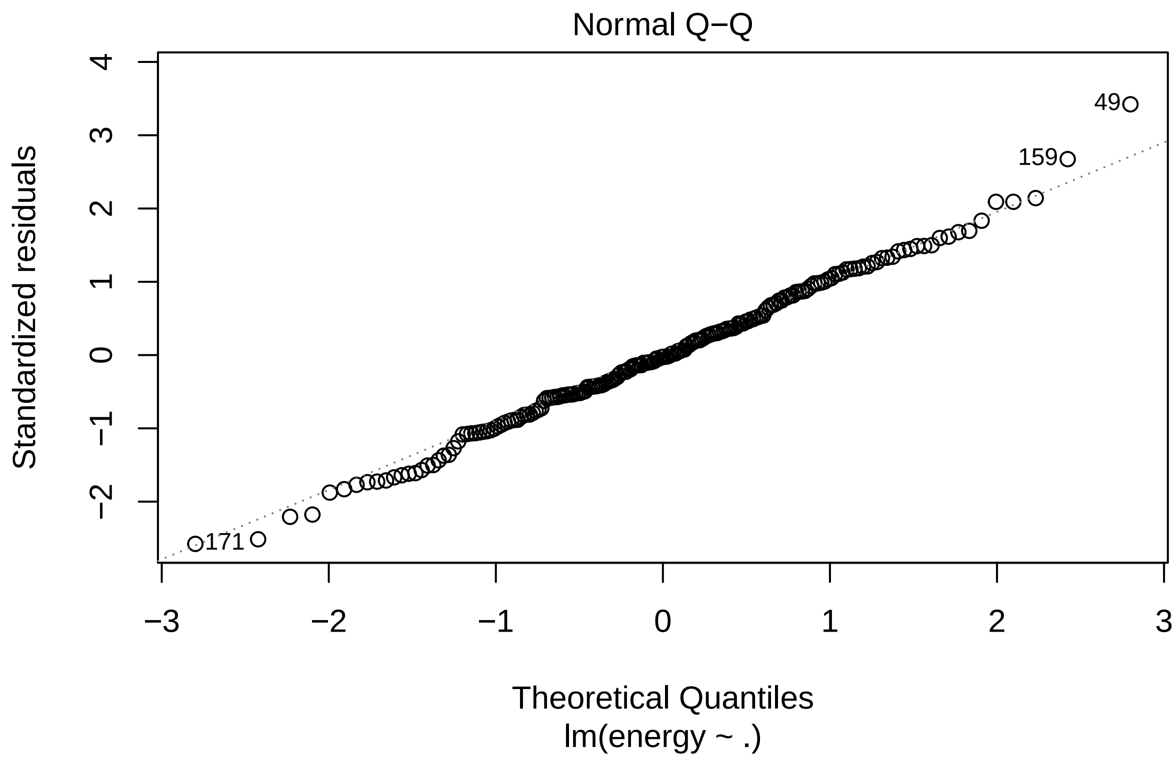
Problem-2 (2 Points)

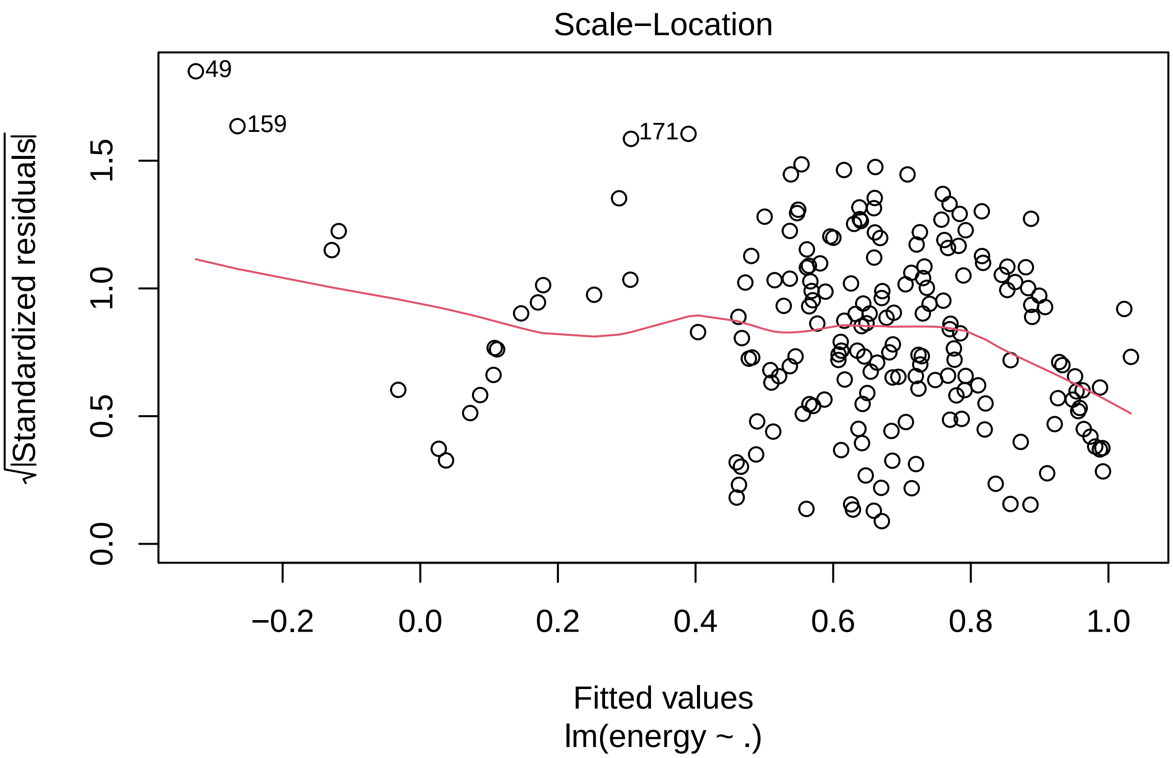
Fit a linear model to predict the *energy* rating using *all* other attributes.Get the summary of the model and explain the results in detail.[*Hint* : Use the lm() function. Click here To get the documentation of the same.]

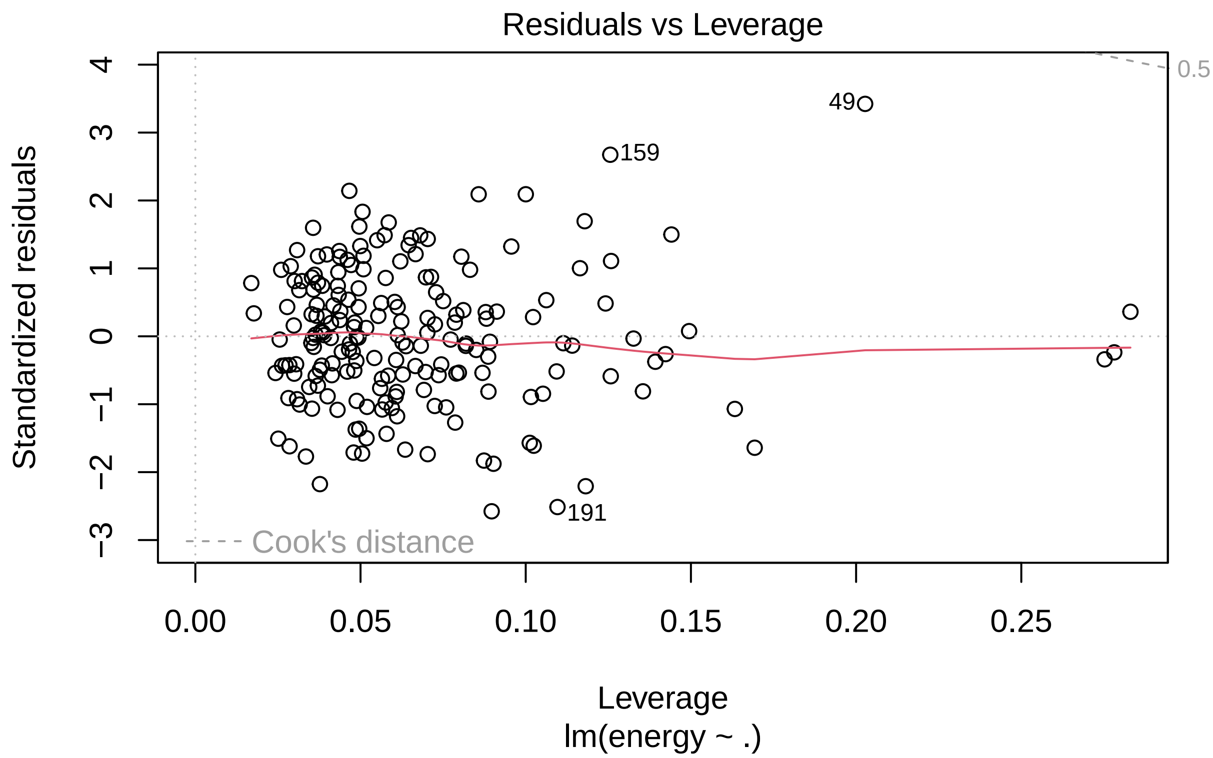
```r
lm_pred <- lm(formula=energy ~ .,data=spotify_df)

plot(lm_pred)
```


Residuals vs Fitted
Residuals
0.3
0.1
−0.1
−0.3
49
159
171
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
Fitted values
lm(energy ~ .)


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(energy ~ .)
49
159
171
−3
−2
−1
0
1
2
3
4


Scale−Location
49
159
171
√|Standardized residuals|
0.0
0.5
1.0
1.5
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
Fitted values
lm(energy ~ .)




```
summary(lm_pred)
```

```
##
## Call:
## lm(formula = energy ~ ., data = spotify_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.26238 -0.05992 -0.00255  0.07276  0.32616
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -0.17513    0.10422  -1.680  0.09459 .
## danceability      -0.27128    0.05478  -4.952 1.67e-06 ***
## key                0.04190    0.02537   1.652  0.10030
## loudness           1.12359    0.07305  15.381  < 2e-16 ***
## mode              -0.02511    0.01589  -1.580  0.11582
## speechiness        0.02627    0.03918   0.670  0.50343
## acousticness      -0.27894    0.03358  -8.306 2.21e-14 ***
## instrumentalness   0.10937    0.04086   2.677  0.00811 **
## liveness           0.02970    0.04594   0.646  0.51880
## valence            0.18905    0.03588   5.269 3.85e-07 ***
## tempo             -0.02676    0.03681  -0.727  0.46817
## duration_ms       -0.03911    0.06926  -0.565  0.57298
## time_signature     0.05589    0.07471   0.748  0.45535
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1067 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
```

Analysis:

The min value of the residuals is -.26070 and the max value is 0.32407. The median is -0.00253. The residual standard error is 0.106 on 182 degrees of freedom. Multiple R-squared value is 0.844. Adjusted R-squared value is 0.8338.P-value is less than 2.2e-16.

Since the F-statistic is 82.08.[high], we reject the Null Hypothesis

Problem-3 (2 points)

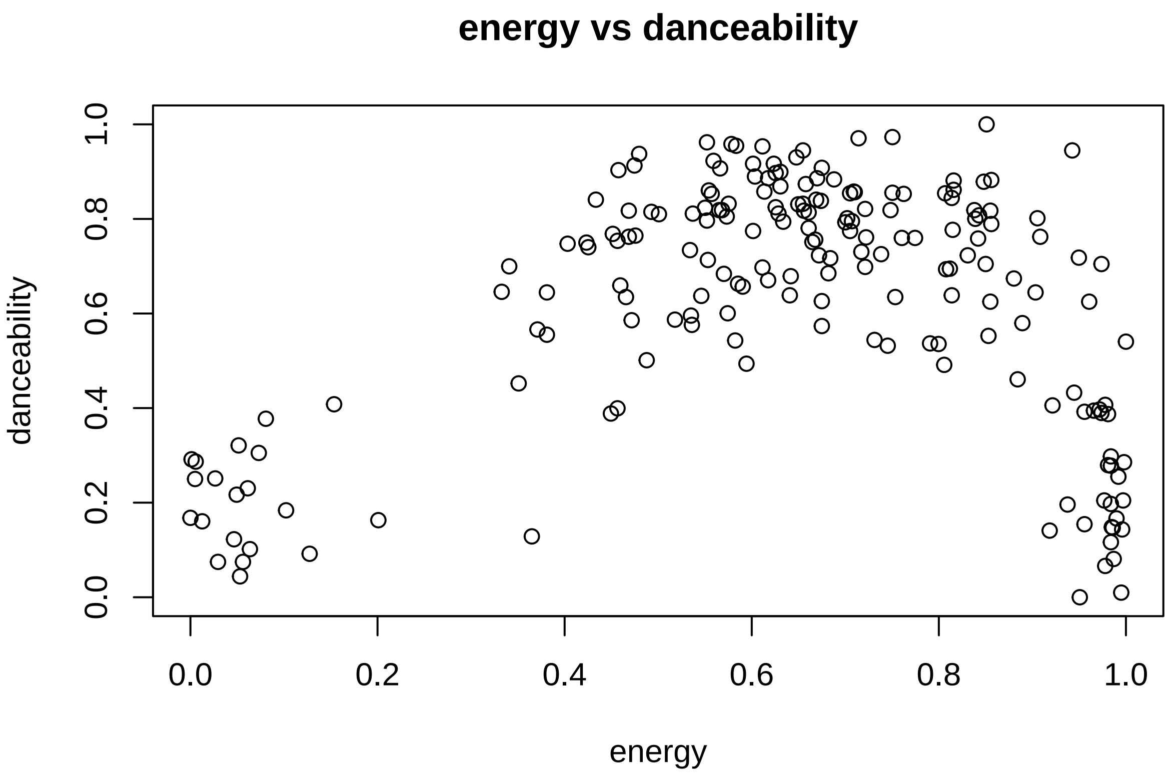
With the help of a correlogram and scatter plots, choose the features you think are important and model an mlr. Justify your choice and explain the new findings.

```r
library(corrplot)
```

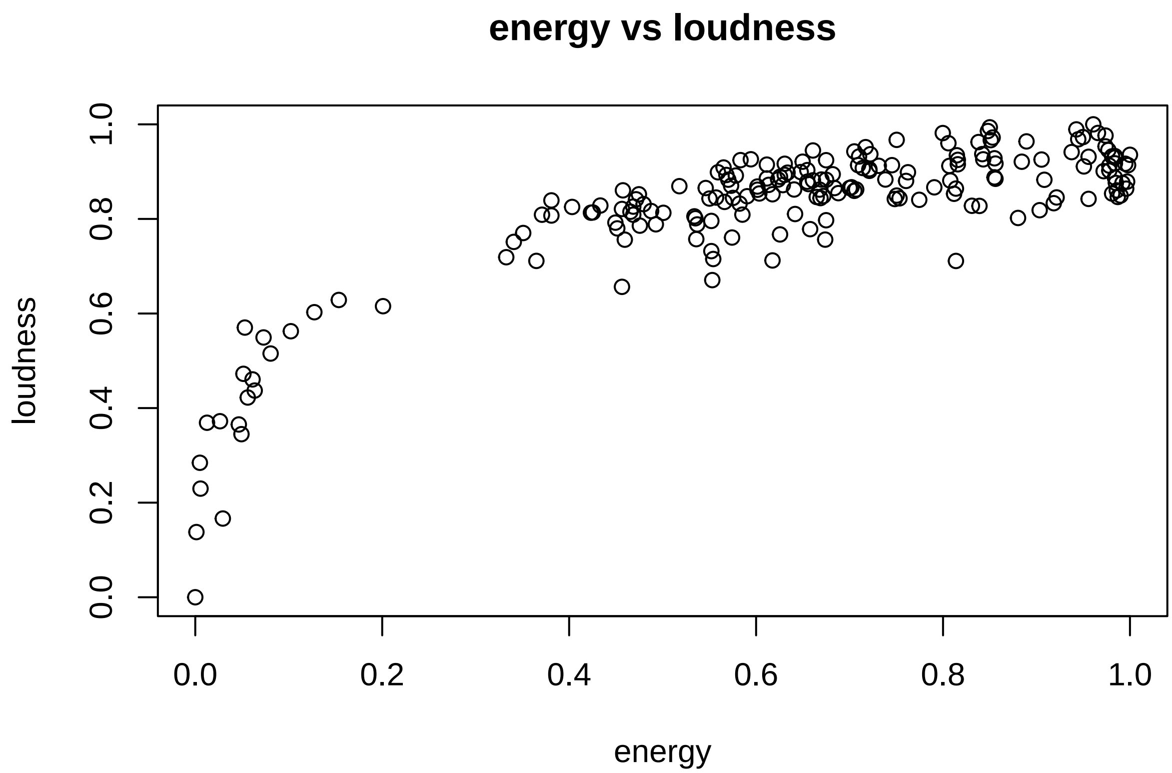
```
## corrplot 0.92 loaded
```

```r
data<-cor(spotify_df)
#finding correlogram
corrplot(data,method="number")
```

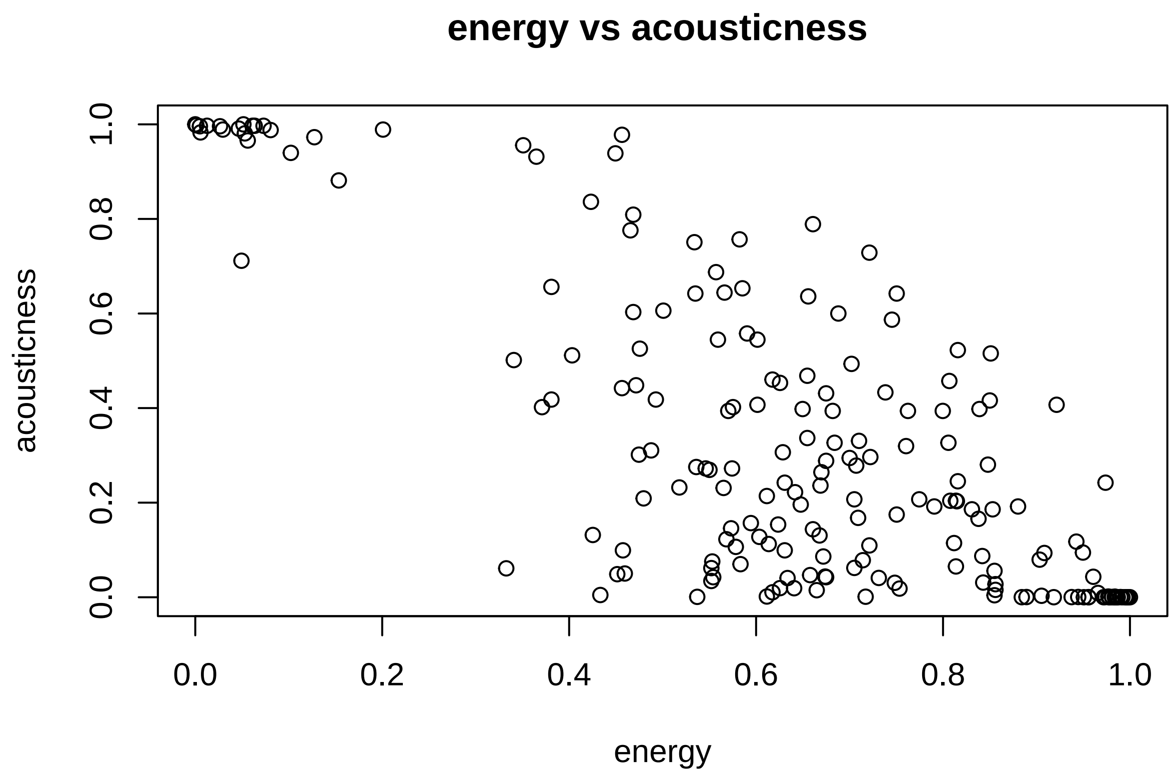

```r
#scatter plots
plot(x=spotify_df$energy , y=spotify_df$danceability,xlab="energy",ylab = "danceability",main="energy v
```



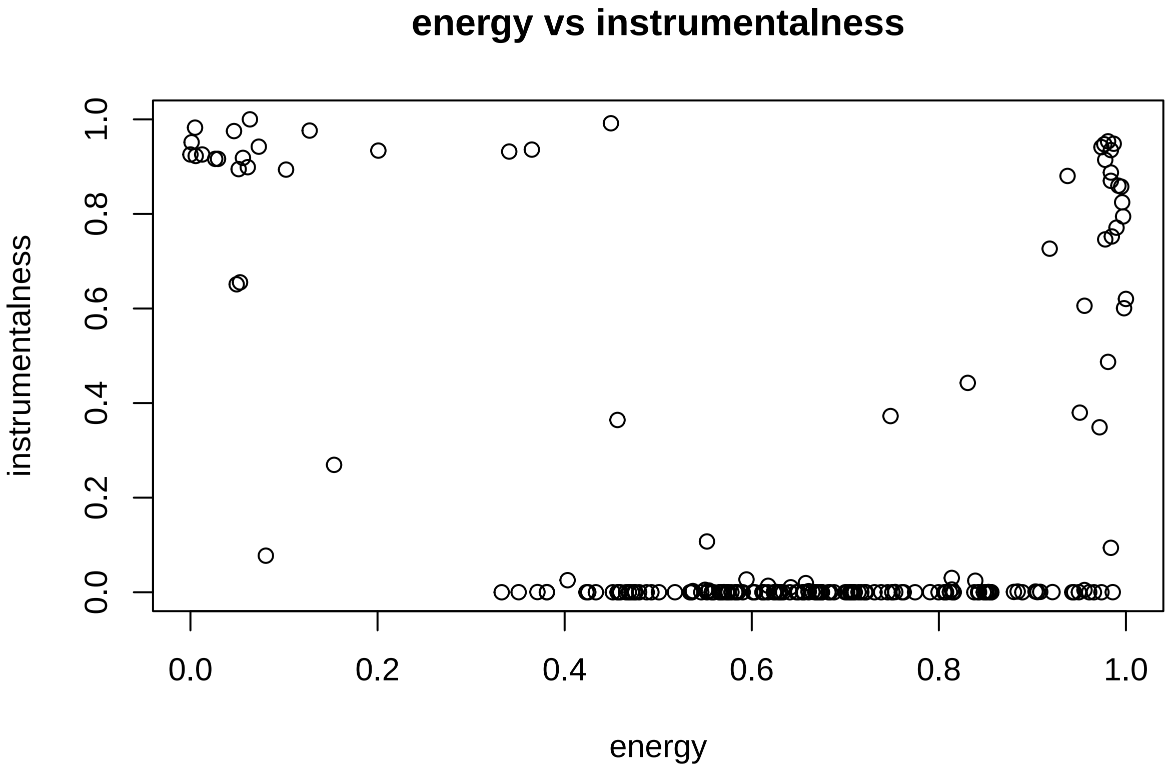

```
plot(x=spotify_df$energy , y=spotify_df$loudness, xlab="energy", ylab="loudness", main="energy vs loudne
```



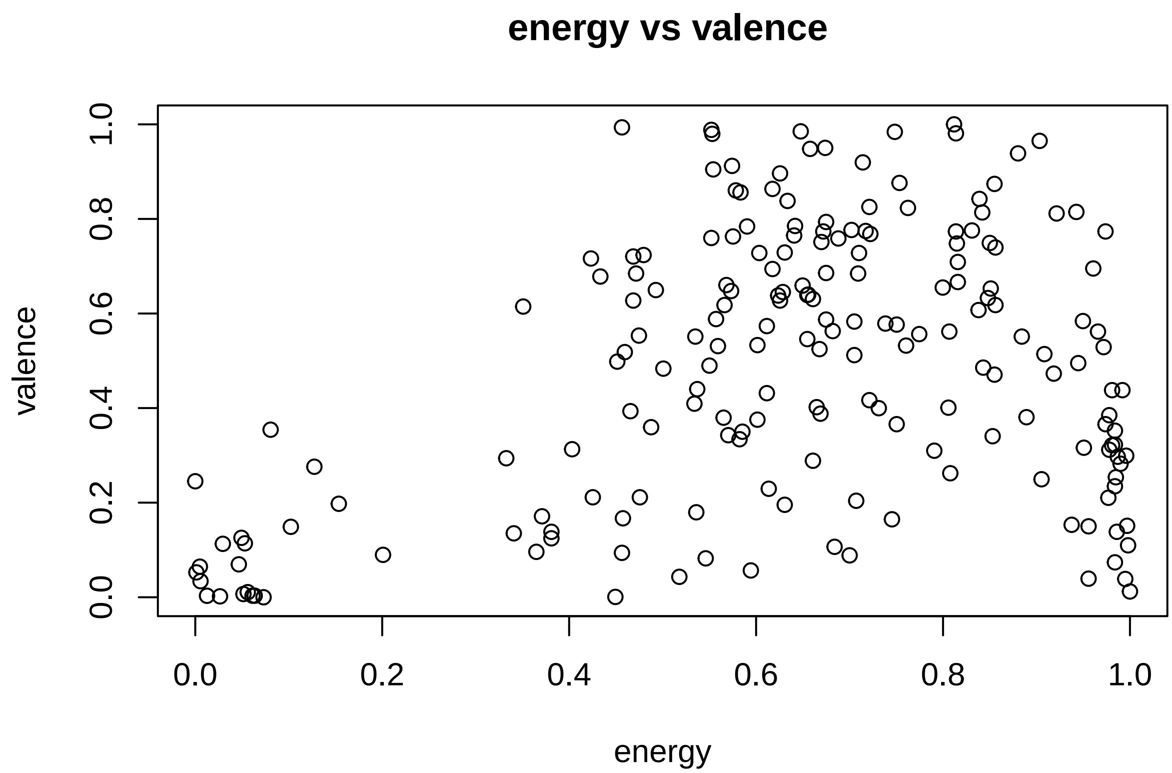

```
plot(x=spotify_df$energy , y=spotify_df$acousticness, xlab="energy", ylab="acousticness", main="energy v
```




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




```
plot(x=spotify_df$energy , y=spotify_df$valence, xlab="energy", ylab="valence", main=" energy vs valenc
```




```
#Finding the Relative Importance of variables
library(relaimpo)
```

```
## Loading required package: MASS

##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select

## Loading required package: boot

##
## Attaching package: 'boot'

## The following object is masked from 'package:lattice':
##
##     melanoma

## Loading required package: survey

## Loading required package: grid

## Loading required package: Matrix

##
## Attaching package: 'Matrix'

## The following objects are masked from 'package:tidyr':
```

```
##
##     expand, pack, unpack

## Loading required package: survival

##
## Attaching package: 'survival'

## The following object is masked from 'package:boot':
##
##     aml

## The following object is masked from 'package:caret':
##
##     cluster

##
## Attaching package: 'survey'

## The following object is masked from 'package:graphics':
##
##     dotchart

## Loading required package: mitools

## This is the global version of package relaimpo.

## If you are a non-US user, a version with the interesting additional metric pmvd is available

## from Ulrike Groempings web site at prof.beuth-hochschule.de/groemping.
```

```
regressor <- lm(energy ~ . , data = spotify_df) # fit lm() model
relImportance <- calc.relimp(regressor, type = "lmg", rela = TRUE)
sort(relImportance$lmg, decreasing=TRUE) # relative importance
```

```
##         loudness     acousticness          valence     danceability
##      0.488260623      0.355156971      0.047325139      0.028356838
## instrumentalness         liveness            tempo              key
##      0.027574268      0.013432729      0.012597091      0.009575742
##      speechiness      duration_ms   time_signature             mode
##      0.005128007      0.004408885      0.004255113      0.003928594
```

```
#From the RelImportance we can find the Importance of each variable and select that variable that contr
reduced_data<-lm(energy ~ danceability+instrumentalness+loudness+acousticness+valence, data=spotify_df)
reduced_data
```

```
##
## Call:
## lm(formula = energy ~ danceability + instrumentalness + loudness +
##     acousticness + valence, data = spotify_df)
##
## Coefficients:
##      (Intercept)      danceability  instrumentalness          loudness
##         -0.13939          -0.28784           0.08986           1.14236
##     acousticness           valence
##         -0.27539           0.18718
```

Reasoning:

After finding the Relative Importance of each column. We selected that Columns that contributed the most.

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_data,lm_pred)
```

```
## Analysis of Variance Table
##
## Model 1: energy ~ danceability + instrumentalness + loudness + acousticness +
##     valence
## 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    189 2.1622
## 2    182 2.0733  7   0.08892 1.1151 0.3554
```

Analysis:

Null Hypothesis: The attributes not chosen are not Significant in predicting the energy value

Alternate Hypothesis : The attributes not chosen are Significant i predicting the energy value.

Since the P-value is 0.3554, which is not less than(alpha=0.05).

we fail to reject the null hypothesis.

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.)

```
library(olsrr)
```

```
##
## Attaching package: 'olsrr'

## The following object is masked from 'package:MASS':
##
##     cement

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

```
aic_stepwise <- lm(energy ~ .,data=spotify_df)
ols_step_both_aic(aic_stepwise)
```

```
##
##
##                                  Stepwise Summary
## ------------------------------------------------------------------------------------
## Variable              Method        AIC         RSS      Sum Sq      R-Sq      Adj. R-Sq
## ------------------------------------------------------------------------------------
## loudness             addition    -175.784     4.495      8.799     0.66189       0.66014
```

```
## acousticness        addition    -239.021    3.217    10.077    0.75803      0.75551
## danceability        addition    -285.619    2.507    10.787    0.81141      0.80844
## valence             addition    -307.057    2.223    11.071    0.83276      0.82924
## instrumentalness    addition    -310.477    2.162    11.131    0.83735      0.83305
## mode                addition    -311.706    2.127    11.167    0.84002      0.83491
## key                 addition    -312.104    2.101    11.193    0.84198      0.83606
## ------------------------------------------------------------------------------------
```

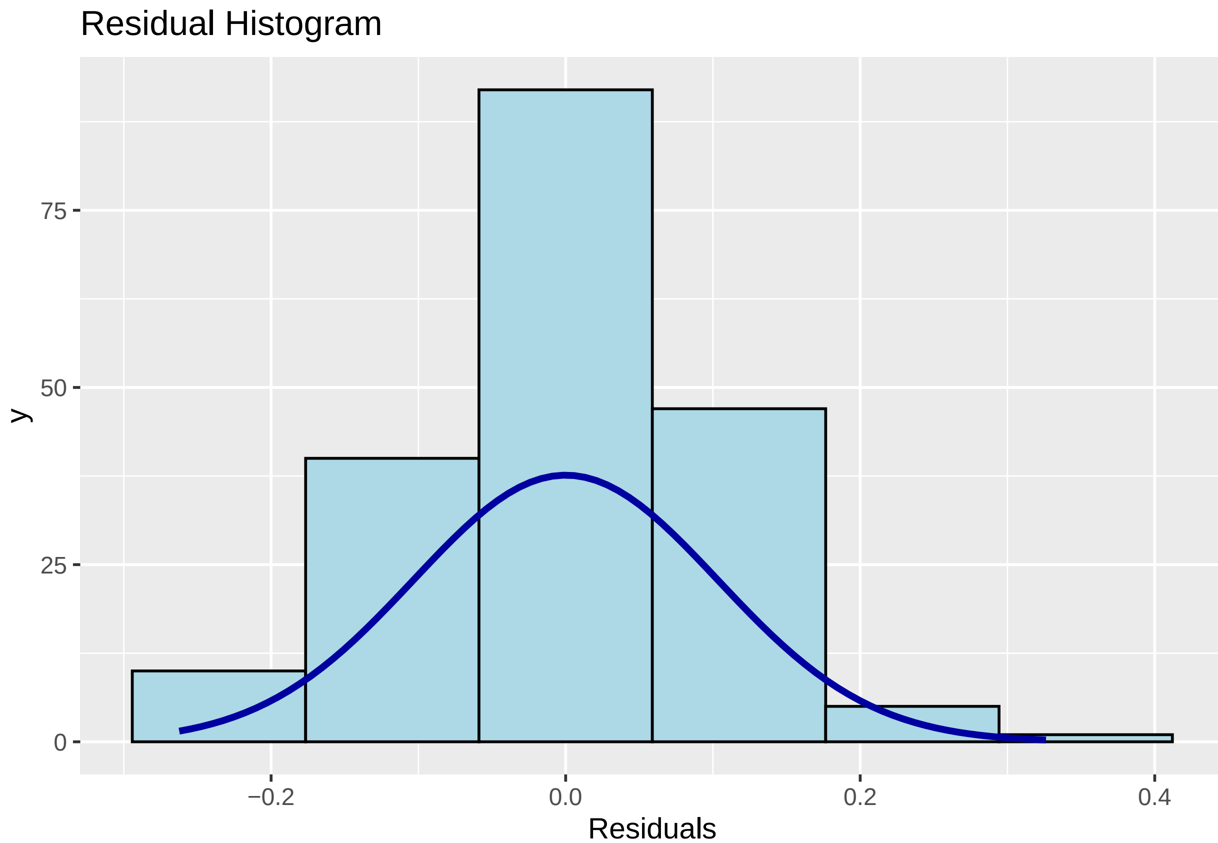
```r
summary(aic_stepwise)
```

```
##
## Call:
## lm(formula = energy ~ ., data = spotify_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.26238 -0.05992 -0.00255  0.07276  0.32616
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -0.17513    0.10422  -1.680  0.09459 .
## danceability     -0.27128    0.05478  -4.952 1.67e-06 ***
## key               0.04190    0.02537   1.652  0.10030
## loudness          1.12359    0.07305  15.381  < 2e-16 ***
## mode             -0.02511    0.01589  -1.580  0.11582
## speechiness       0.02627    0.03918   0.670  0.50343
## acousticness     -0.27894    0.03358  -8.306 2.21e-14 ***
## instrumentalness  0.10937    0.04086   2.677  0.00811 **
## liveness          0.02970    0.04594   0.646  0.51880
## valence           0.18905    0.03588   5.269 3.85e-07 ***
## tempo            -0.02676    0.03681  -0.727  0.46817
## duration_ms      -0.03911    0.06926  -0.565  0.57298
## time_signature    0.05589    0.07471   0.748  0.45535
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1067 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
```

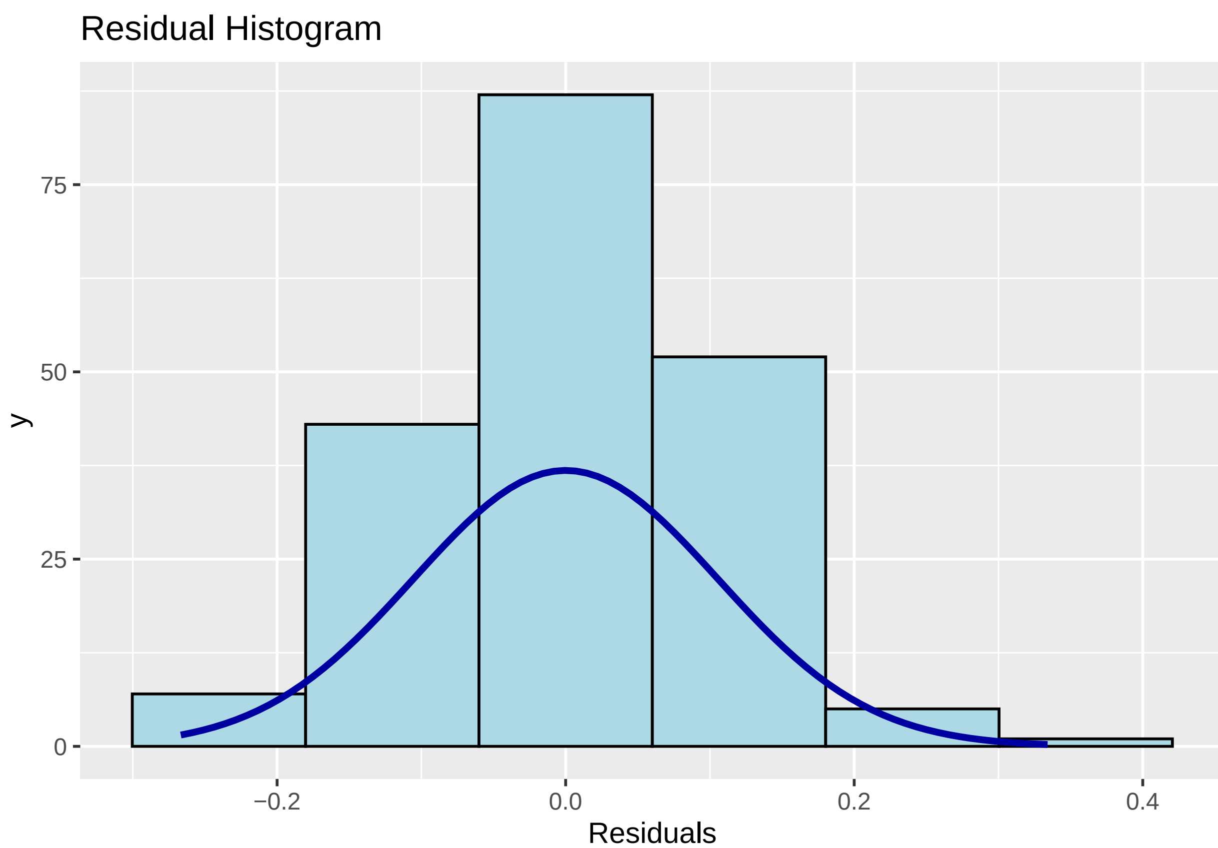
Problem-6 (1 Point)

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

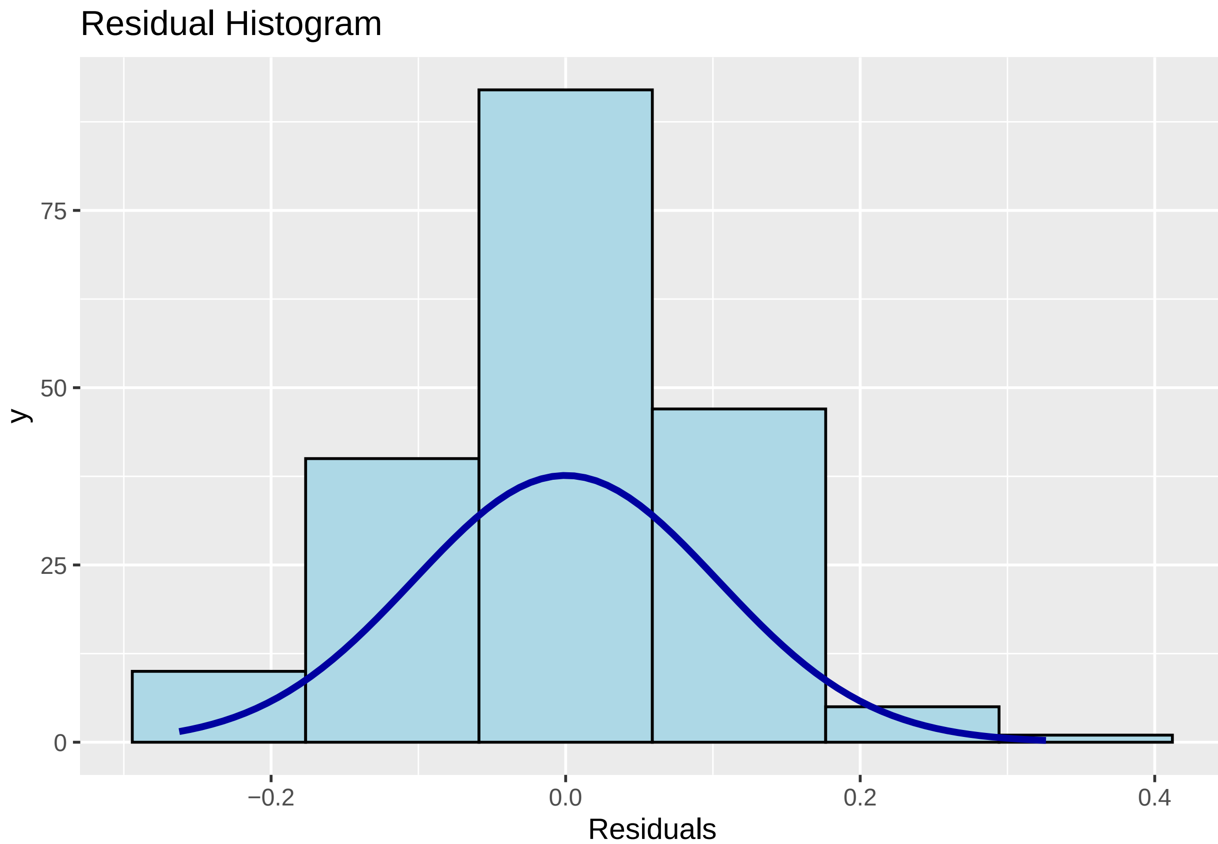
```r
ols_plot_resid_hist(lm_pred)
```



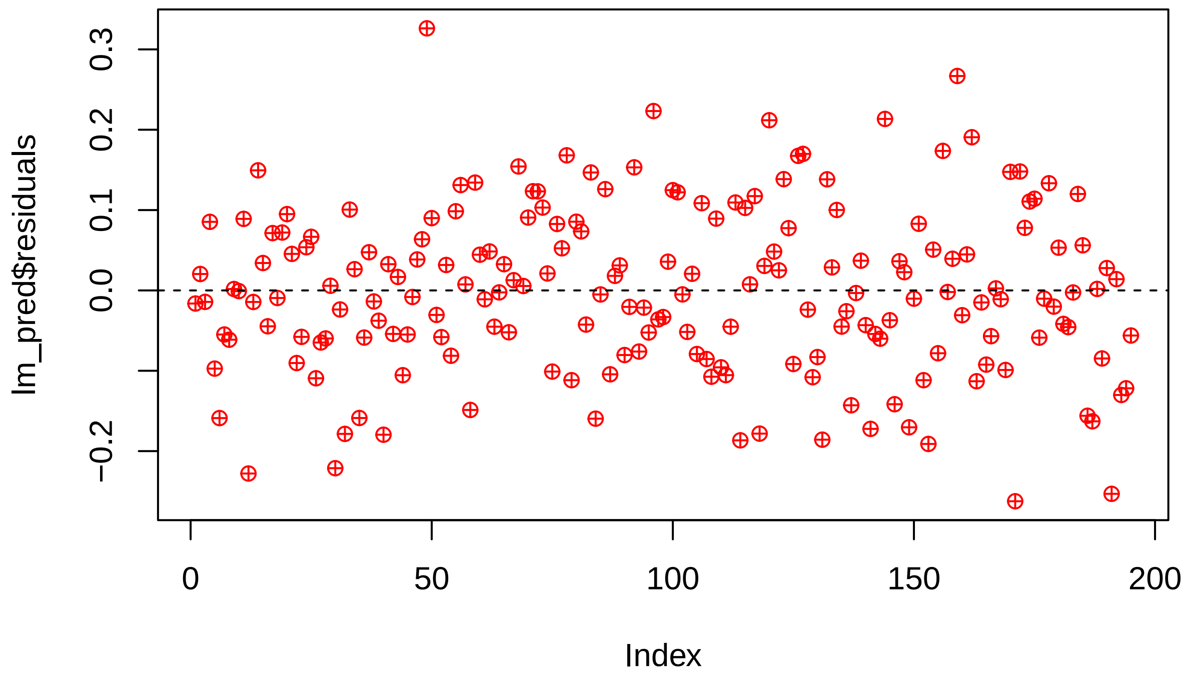

```
ols_plot_resid_hist(reduced_data)
```



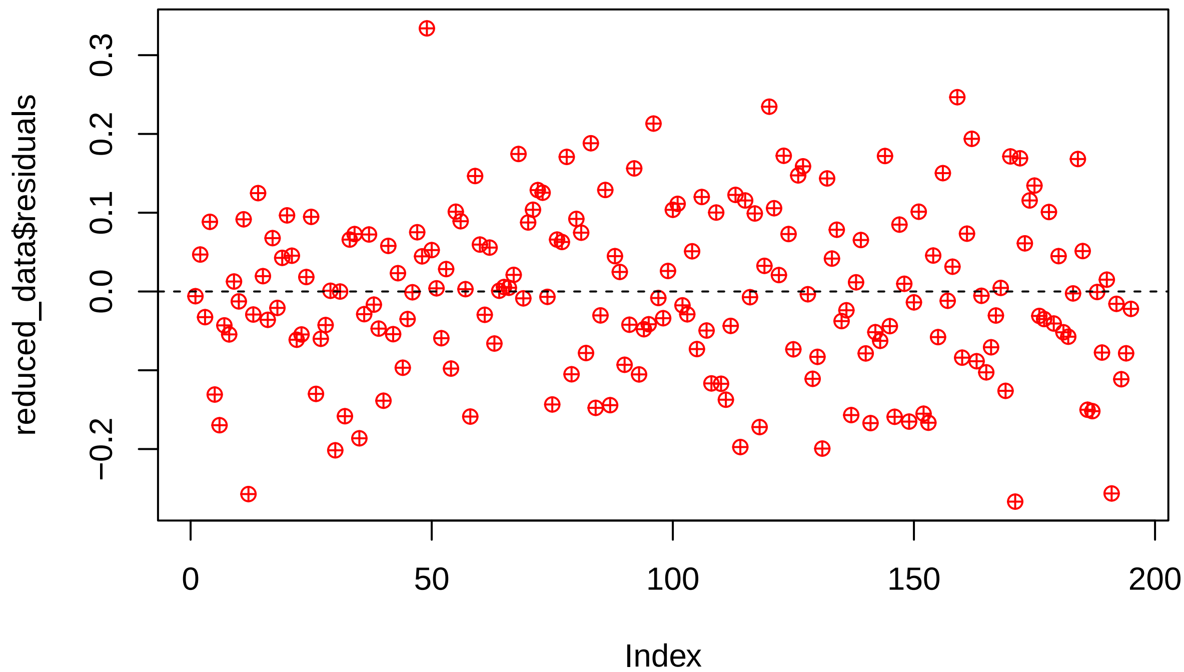

```
ols_plot_resid_hist(aic_stepwise)
```



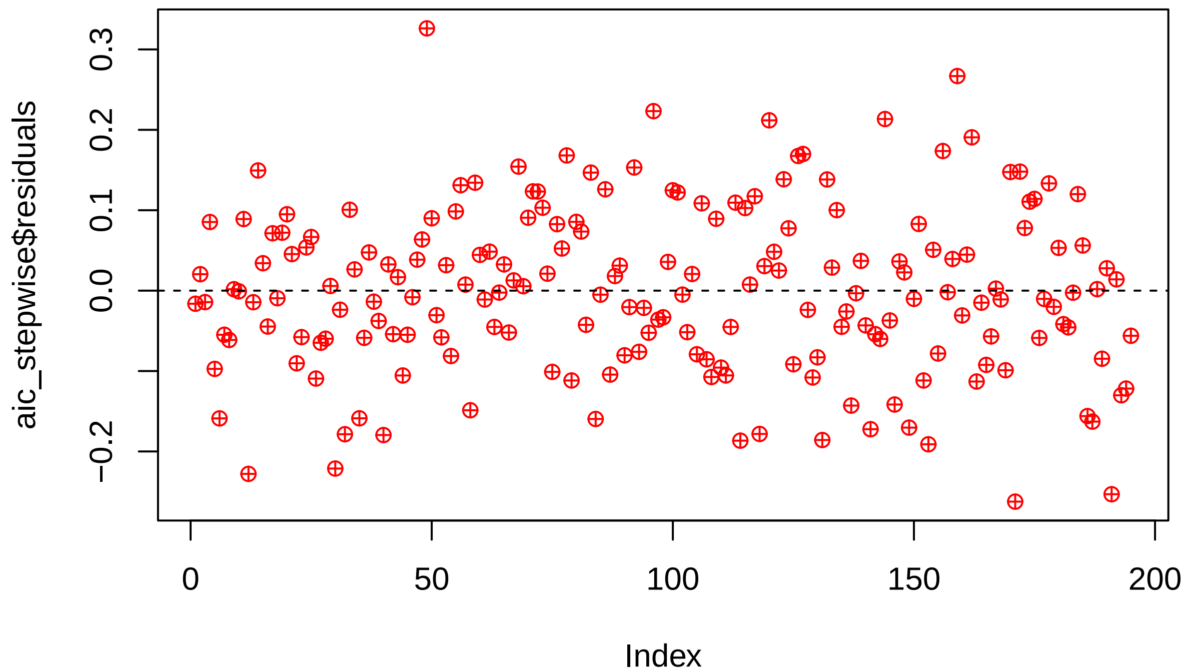

```
plot(lm_pred$residuals,pch =10 ,col="red")
abline(h=0,lty=2)
```




```
plot(reduced_data$residuals,pch =10 ,col="red")
abline(h=0,lty=2)
```




```
plot(aic_stepwise$residuals,pch =10 ,col="red")
abline(h=0,lty=2)
```



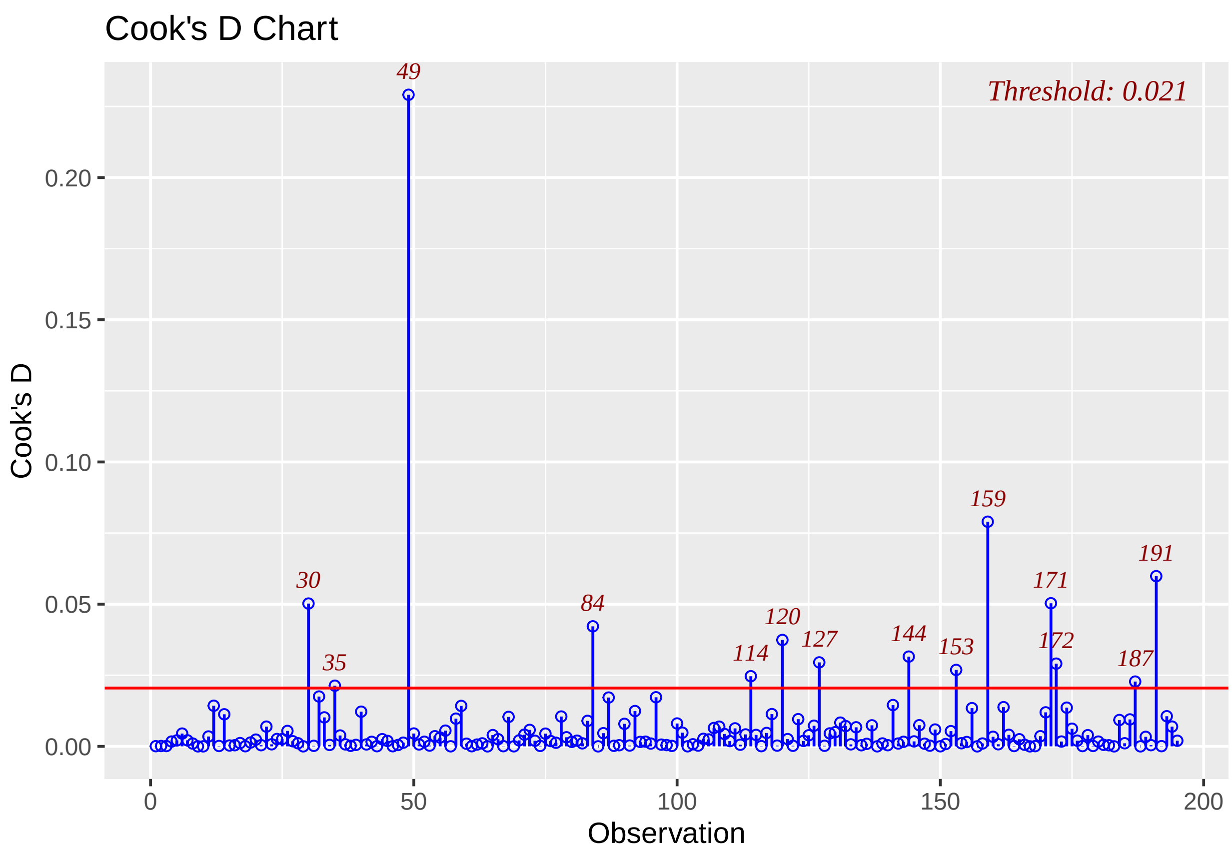
Problem-7 (2 Points)

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.]

```
#detemining multicollinearity using VIF
ols_vif_tol(lm_pred)
```

```
##           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
```

```
#representing Outliers
cookd <- ols_plot_cooksd_chart(lm_pred);
```




```
#Yes there are outliers in the dataset.

#Removing Outliers
fit<- lm(energy ~ ., data = spotify_df);
spotify_df$cooksd <- cooks.distance(fit); # Defining outliers based on 4/n criteria
#Creating a new column called outlier which tells if that row of data is an outlier or not.
spotify_df$outlier <- ifelse(spotify_df$cooksd < 4/nrow(spotify_df), "keep","delete");
#Now we are trying to delete all the rows that contain the data "outlier = delete "
spotify_df=spotify_df[!grepl("delete", spotify_df$outlier),];
head(spotify_df)
```

```
##   danceability     energy        key  loudness mode speechiness acousticness
## 1    0.8247549 0.62560386 0.63636364 0.8890920    0  0.03885201   0.45326466
## 2    0.7745098 0.70511272 0.90909091 0.8593613    0  0.54314721   0.20703275
## 3    0.1605392 0.01258052 0.09090909 0.3690169    1  0.02752831   0.99698492
## 4    0.7254902 0.73832528 0.27272727 0.8833312    0  0.05993752   0.43316409
## 5    0.8051471 0.57326892 0.09090909 0.8702567    1  0.37914877   0.14572602
## 6    0.7941176 0.63365539 0.72727273 0.8978334    1  0.18976962   0.04060007
##   instrumentalness   liveness     valence     tempo duration_ms time_signature
## 1     7.574819e-04 0.11151859 0.627394940 0.2986443   0.3932821           0.75
## 2     0.000000e+00 0.09684947 0.512014396 0.7605056   0.2940693           0.75
## 3     9.256966e-01 0.11485248 0.003069758 0.1261836   0.3629418           0.75
## 4     1.217750e-06 0.14985831 0.578702234 0.2476870   0.2278801           0.75
## 5     0.000000e+00 0.07034506 0.647507145 0.7921078   0.1768309           0.75
## 6     0.000000e+00 0.09684947 0.838043823 0.6739248   0.2540198           0.75
##         cooksd outlier
```

```
## 1 6.922961e-05    keep
## 2 1.258874e-04    keep
## 3 1.080694e-04    keep
## 4 1.695989e-03    keep
## 5 2.096273e-03    keep
## 6 4.502237e-03    keep
```

```
#Now this spotify_df is free of outliers..


summary(spotify_df)
```

```
##   danceability       energy              key            loudness
##  Min.   :0.0000   Min.   :0.004952   Min.   :0.0000   Min.   :0.2297
##  1st Qu.:0.4914   1st Qu.:0.552134   1st Qu.:0.1818   1st Qu.:0.8137
##  Median :0.7230   Median :0.669887   Median :0.5455   Median :0.8668
##  Mean   :0.6385   Mean   :0.660475   Mean   :0.5073   Mean   :0.8363
##  3rd Qu.:0.8248   3rd Qu.:0.842995   3rd Qu.:0.7273   3rd Qu.:0.9137
##  Max.   :1.0000   Max.   :1.000000   Max.   :1.0000   Max.   :1.0000
##       mode         speechiness       acousticness    instrumentalness
##  Min.   :0.0000   Min.   :0.00000   Min.   :0.0000   Min.   :0.0000000
##  1st Qu.:0.0000   1st Qu.:0.05857   1st Qu.:0.0406   1st Qu.:0.0000000
##  Median :1.0000   Median :0.14096   Median :0.2070   Median :0.0000043
##  Mean   :0.5359   Mean   :0.23576   Mean   :0.3001   Mean   :0.1791565
##  3rd Qu.:1.0000   3rd Qu.:0.39672   3rd Qu.:0.4573   3rd Qu.:0.0254902
##  Max.   :1.0000   Max.   :1.00000   Max.   :1.0000   Max.   :1.0000000
##     liveness          valence            tempo          duration_ms
##  Min.   :0.00000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.08235   1st Qu.:0.2887   1st Qu.:0.3408   1st Qu.:0.1768
##  Median :0.11985   Median :0.5311   Median :0.5245   Median :0.2230
##  Mean   :0.18734   Mean   :0.4989   Mean   :0.5104   Mean   :0.2380
##  3rd Qu.:0.23654   3rd Qu.:0.7280   3rd Qu.:0.6901   3rd Qu.:0.2851
##  Max.   :1.00000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##  time_signature      cooksd             outlier
##  Min.   :0.000   Min.   :2.490e-07   Length:181
##  1st Qu.:0.750   1st Qu.:4.061e-04   Class :character
##  Median :0.750   Median :1.675e-03   Mode  :character
##  Mean   :0.732   Mean   :3.365e-03
##  3rd Qu.:0.750   3rd Qu.:4.634e-03
##  Max.   :1.000   Max.   :1.752e-02
```