STAT501 Project Report: Million Song Data

Kinnri Sinha Movina Moses Shanmukh Srinivas
Shubham Shetty Sruthi Srinivasan

Introduction

The aim of this project is to perform exploratory data analysis (EDA) on the Million Song Dataset (http://millionsongdataset.com). The Million Song Dataset contains metadata regarding a million popular contemporary English songs released in the 20th and early 21st century. The goal of EDA is to find some interesting trends in the songs dataset which is our basic research question. In this we further analyzing three main topics. First, relation between Artist familiarity and Artist hotness. Second, relation between the artist hotness and the song hotness based on the year of song release. Third, the trends of artist name lengths over the years.

Data

We will be performing our analysis on a subset of the Million Song dataset which is provided by official website. This subset consists of 10,000 songs (1%, 1.8 GB) which are randomly sampled from the larger dataset. Principally, the dataset consists of both metadata and audio analysis features. Each file is for one track which corresponds to one song, one release and one artist. All the information about these four items (track, song, release, artist) are in every file (which involves some redundancy, although the bulk of the data, relating to the audio analysis, is unique).

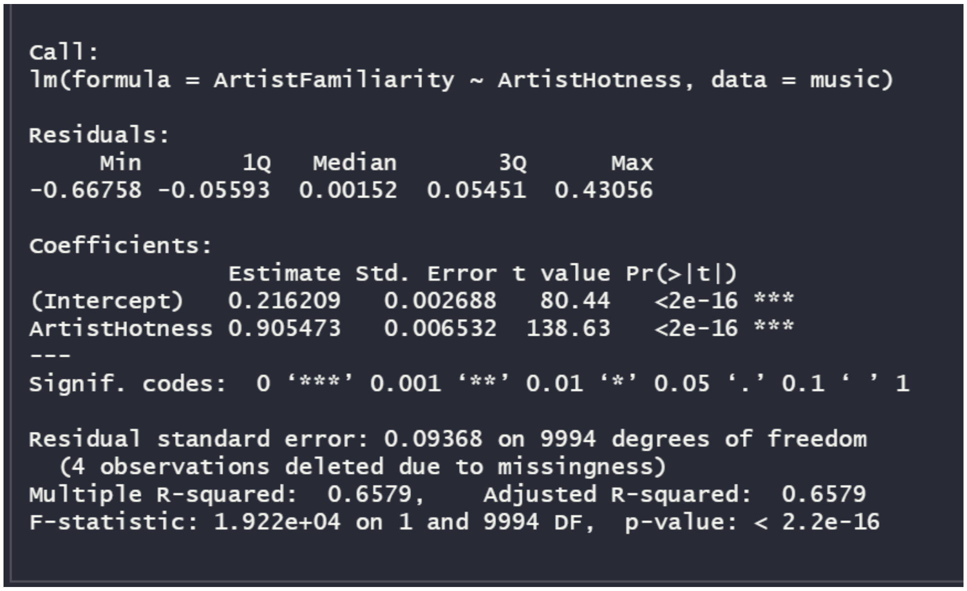
Code

We performed the analysis using R and our code is presented in this report.

Analysis

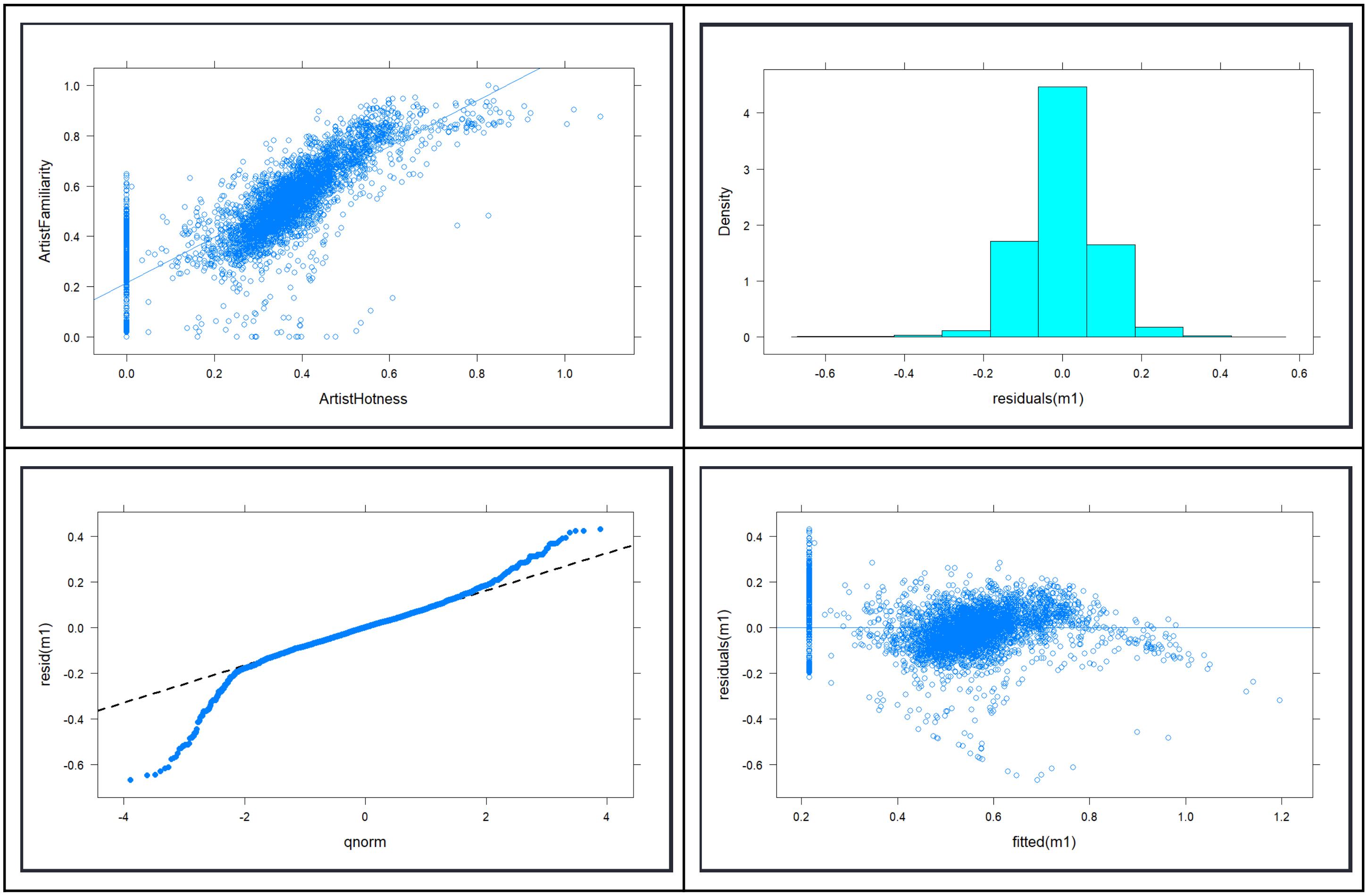
Part 1 - Relationship between Artist Familiarity and Artist Hotness

Here, we fit a model predicting Artist Familiarity using Artist Hotness. It has an R^2 value of 0.5509 indicating that this could be a fair model to use.



```
Call:
lm(formula = ArtistFamiliarity ~ ArtistHotness, data = music)

Residuals:
     Min       1Q   Median       3Q      Max
-0.66758 -0.05593  0.00152  0.05451  0.43056

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.216209   0.002688   80.44   <2e-16 ***
ArtistHotness 0.905473   0.006532  138.63   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.09368 on 9994 degrees of freedom
  (4 observations deleted due to missingness)
Multiple R-squared:  0.6579,    Adjusted R-squared:  0.6579
F-statistic: 1.922e+04 on 1 and 9994 DF,  p-value: < 2.2e-16
```

Checking Assumptions



From the above plots, we checked the following assumptions -

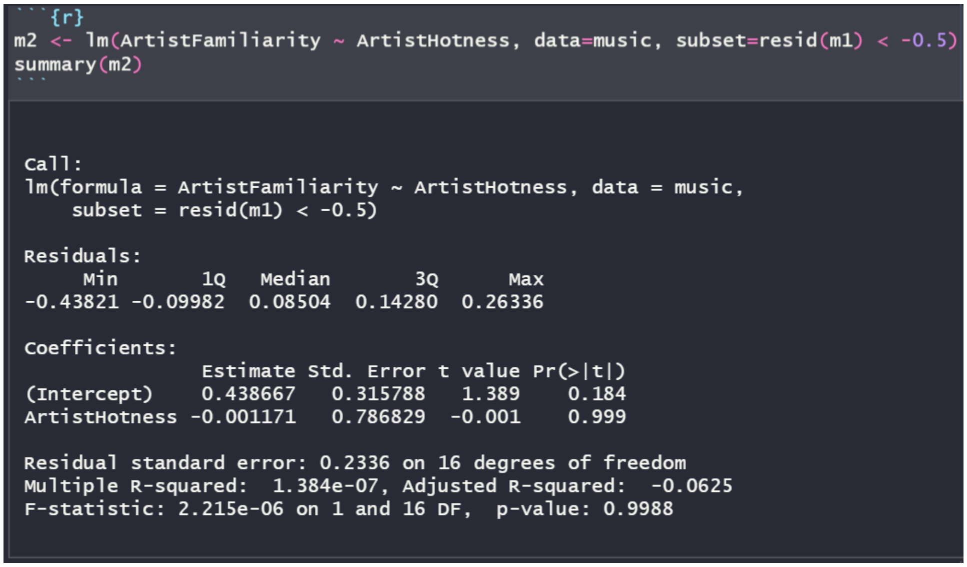
1. Linearity: The linearity assumption is met because the scatter plot does not show any evidence of nonlinearity. It is seen that with an increase in artist familiarity, there is an increase in artist hotness.

2. Independence: The independence assumption appears to be met because the residuals plot does not show any regular pattern.
3. Normal Population: It does appear that the normality assumption is met because in the histogram, we see that the residuals do follow a fairly normal distribution.

4. Equal Variance: The equal variance assumption appears to be met. The residual graph displays an equal variance among all the data. The residuals look fairly evenly scattered about the horizontal line of 0. But there are a very few values that lie below -0.5, these data values could be contributing the slight left-skewedness.

Dropping the Outliers

To see if we would get a better model by dropping the outliers, we experimented by dropping the values with residuals less than -0.5.

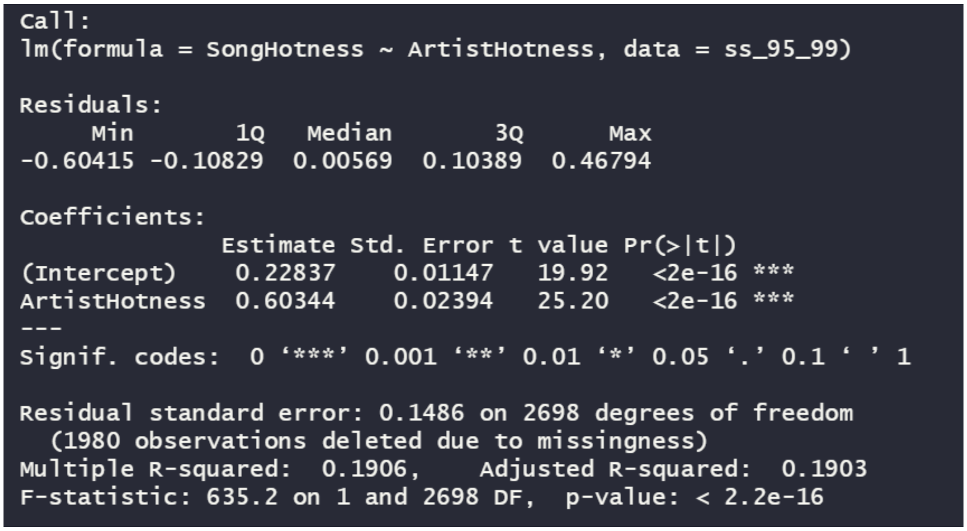


~~~
```{r}
m2 <- lm(ArtistFamiliarity ~ ArtistHotness, data=music, subset=resid(m1) < -0.5)
summary(m2)
```


Call:
lm(formula = ArtistFamiliarity ~ ArtistHotness, data = music,
    subset = resid(m1) < -0.5)

Residuals:
     Min       1Q   Median       3Q      Max
-0.43821 -0.09982  0.08504  0.14280  0.26336

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.438667   0.315788   1.389    0.184
ArtistHotness -0.001171   0.786829  -0.001    0.999

Residual standard error: 0.2336 on 16 degrees of freedom
Multiple R-squared:  1.384e-07, Adjusted R-squared:  -0.0625
F-statistic: 2.215e-06 on 1 and 16 DF,  p-value: 0.9988
~~~

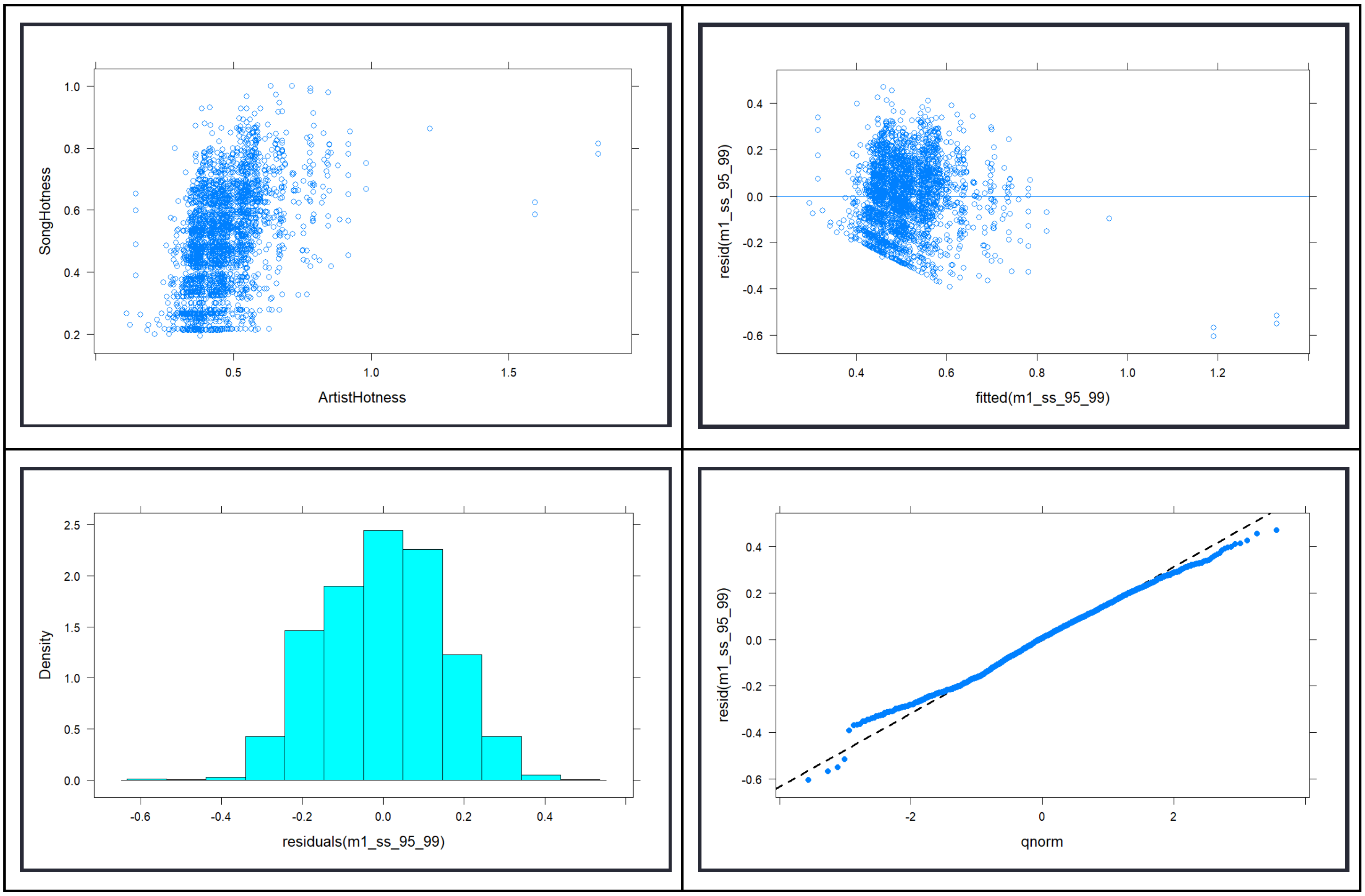
Dropping the residuals resulted in a higher R^2 value, indicating that the quality of the model does increase by dropping the residuals.

Part 2 - A study between the artist hotness and the song hotness based on the year of song release

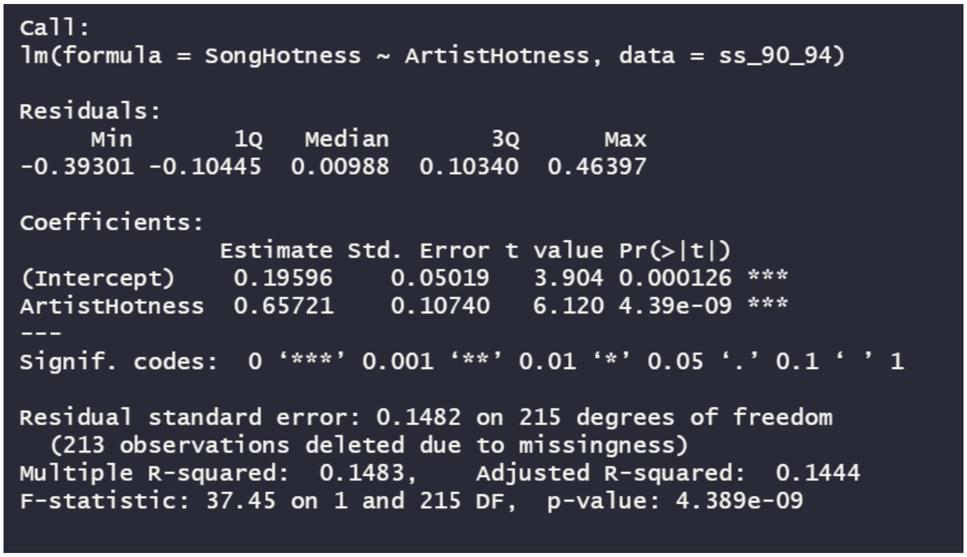
1. 1995-1999: We can see a general increase in Song Hotness with Artist Hotness but the relationship is not significant enough for it to be modeled as a linear relationship. The R-squared value of 20% further indicates that a linear model is not suitable for modeling this relationship. The error plots reveal constant variance, zero mean and an approximately normal distribution.



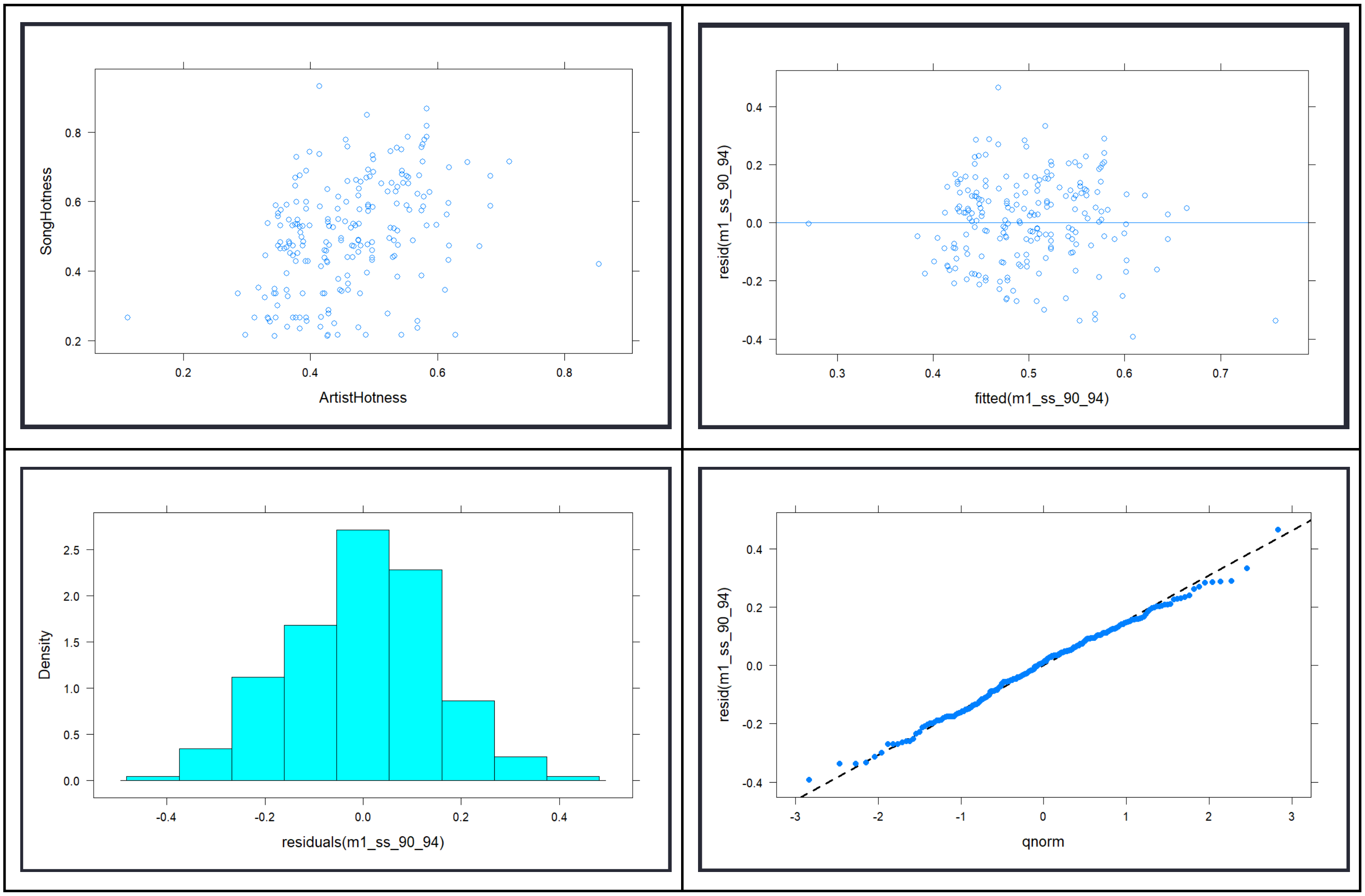
```
Call:
lm(formula = SongHotness ~ ArtistHotness, data = ss_95_99)

Residuals:
     Min       1Q   Median       3Q      Max
-0.60415 -0.10829  0.00569  0.10389  0.46794

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.22837    0.01147   19.92   <2e-16 ***
ArtistHotness  0.60344    0.02394   25.20   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1486 on 2698 degrees of freedom
  (1980 observations deleted due to missingness)
Multiple R-squared:  0.1906,    Adjusted R-squared:  0.1903
F-statistic: 635.2 on 1 and 2698 DF,  p-value: < 2.2e-16
```



2. 1990 - 1994: We do not see a distinctly linear relationship between artist and song hotness. The R-squared value of 15% further indicates that a linear model is not suitable for modeling this relationship. The error plots reveal constant variance, zero mean and an approximately normal distribution.



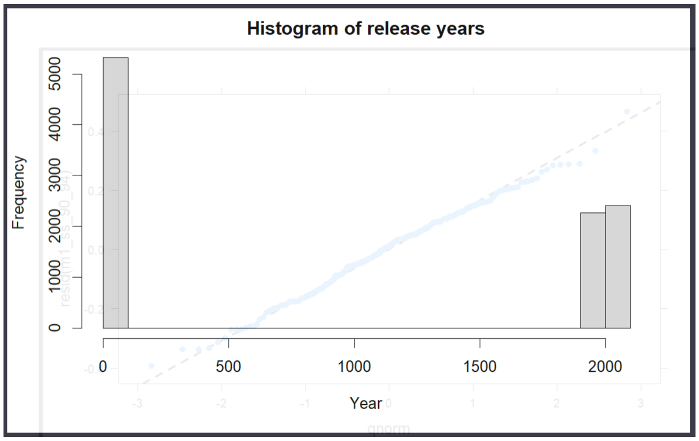
```
Call:
lm(formula = SongHotness ~ ArtistHotness, data = ss_90_94)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.39301 -0.10445  0.00988  0.10340  0.46397 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)    0.19596    0.05019   3.904 0.000126 ***
ArtistHotness  0.65721    0.10740   6.120 4.39e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1482 on 215 degrees of freedom
  (213 observations deleted due to missingness)
Multiple R-squared:  0.1483,	Adjusted R-squared:  0.1444 
F-statistic: 37.45 on 1 and 215 DF,  p-value: 4.389e-09
```

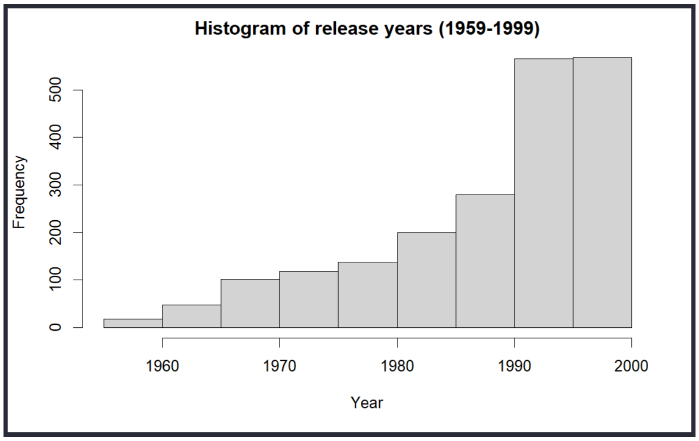


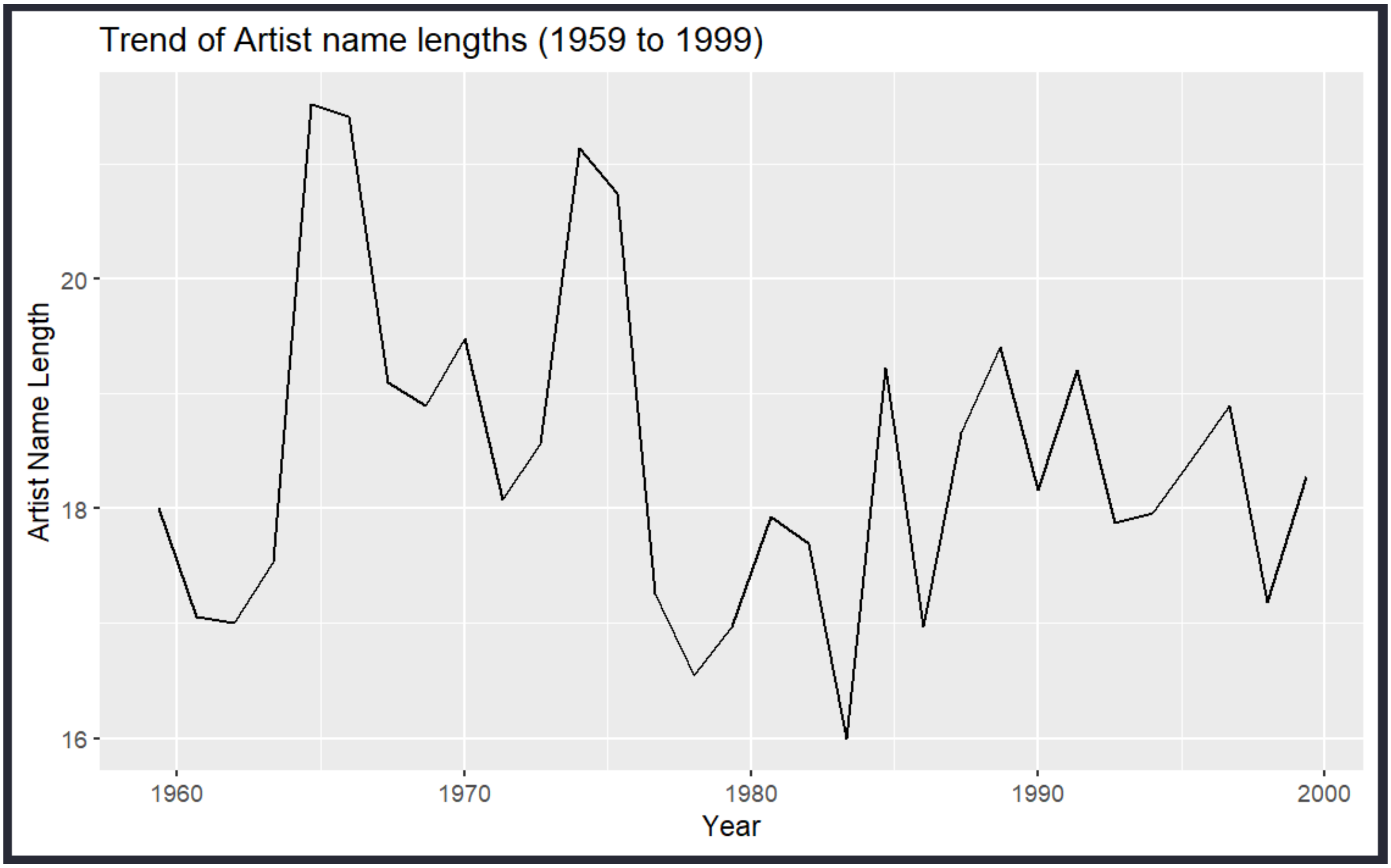
Note: Year-wise analysis for remaining years can be found in the appendix as it produced similar results as other year windows.

Part 3 - A study to analyze the trends of artist name lengths over the years

To find out the trend, we can first plot the histogram of the release years of the songs to get an idea of the range of years to work with. From the histogram of years, we see that there are songs that don't have a year associated with it and hence is zero. We can filter those data points out and generate a new histogram.







From the plot we can see various trends in the length of the Artist names. It seems to have reached an all time high in the mid-1960's and except for a few years in mid 1970's. After the mid 1970's there was a sharp decrease in the length, after which the average name length stayed 15-16 characters till 2000.

This state of constancy is kind of counter-intuitive since we might expect the names to get longer because of a shortage of unique shorter names.

Conclusion

Our basic research question was to analyse the trends in the Million Song DataSet.

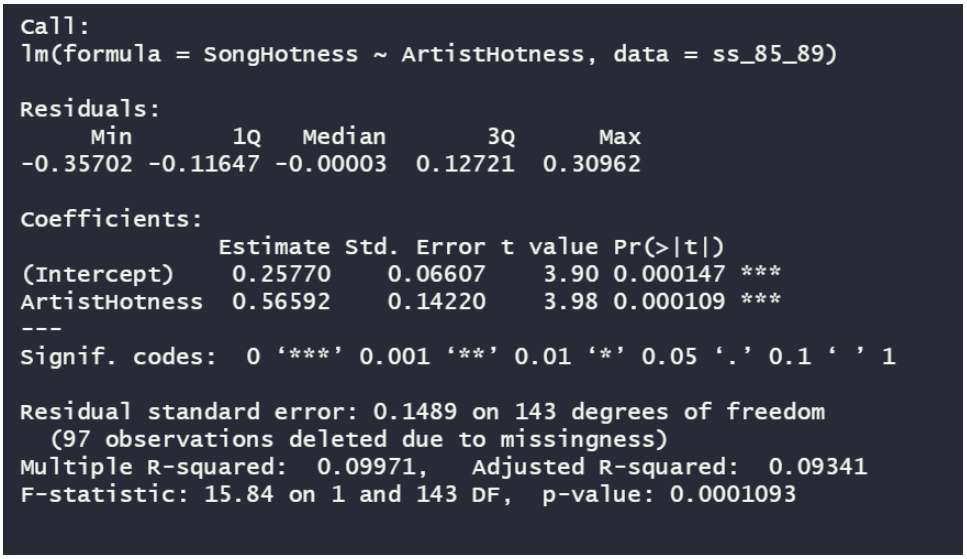
In this we first analysed the relation between Artist familiarity and Artist hotness. We can see that with the increase in Artist familiarity there is an increase in the artist hotness and we can hence say that it is easy to model the artist familiarity to predict the artist hotness.
We also found that the residuals follow a normal distribution.

Second, we analysed the relation between artist hotness and the song hotness based on the year in the year in which the song was released. In general there was increase in Song Hotness with Artist Hotness. This primarily seen between the years 1960 and 1964. Later, it wasn't very distinctly visible.

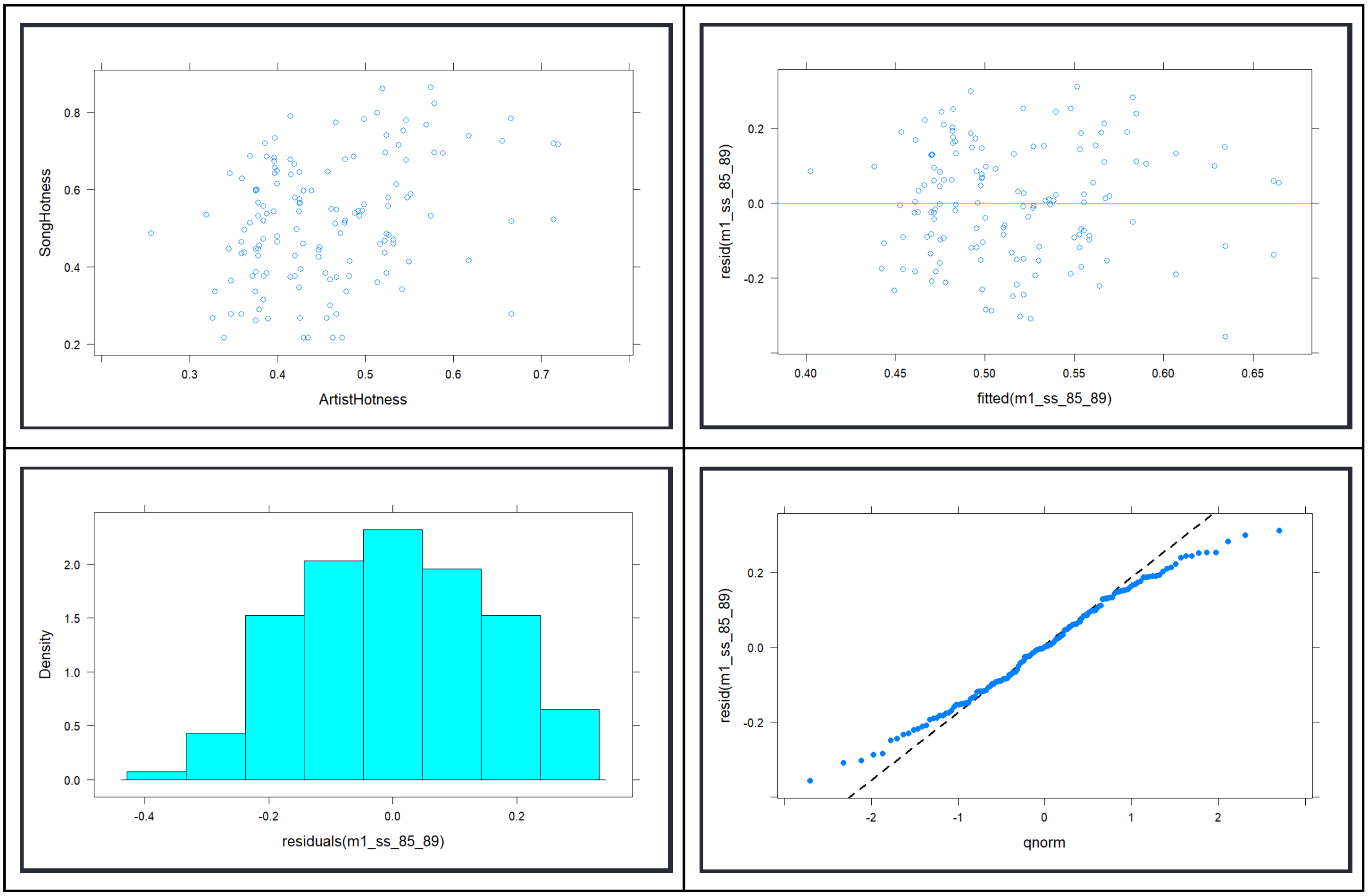
Third, we analysed the trends of artist name lengths over the years.We see that the Artist name length peaked in the mid-1960's, after which there was a sharp decrease in the name length and the name length has remained 15-16 characters till 2000s.

Appendix

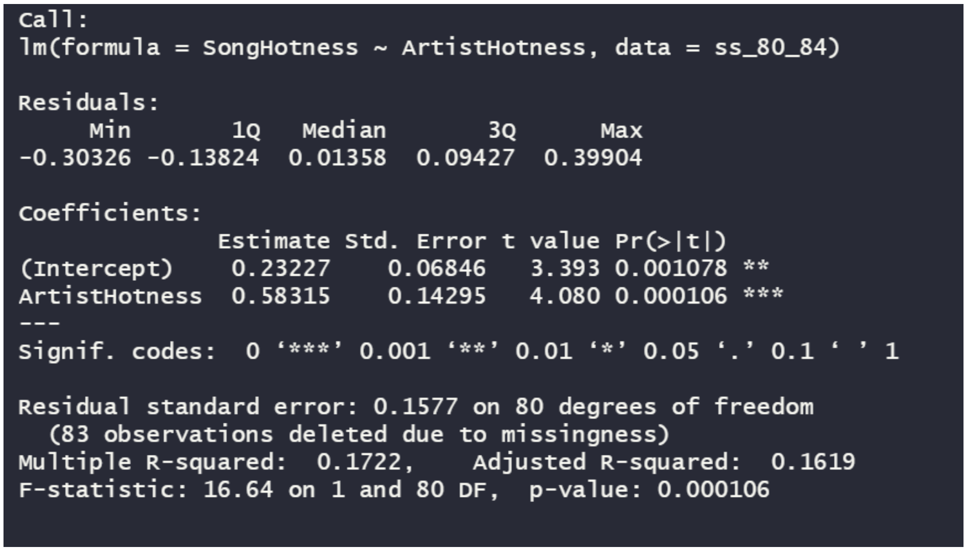
1. 1985 - 1989: We do not see a distinctly linear relationship between artist and song hotness. The R-squared value of 15% further indicates that a linear model is not suitable for modeling this relationship. The error plots reveal constant variance, zero mean, however the curved tails in the normal probability plot indicates a deviation from the normal distribution.



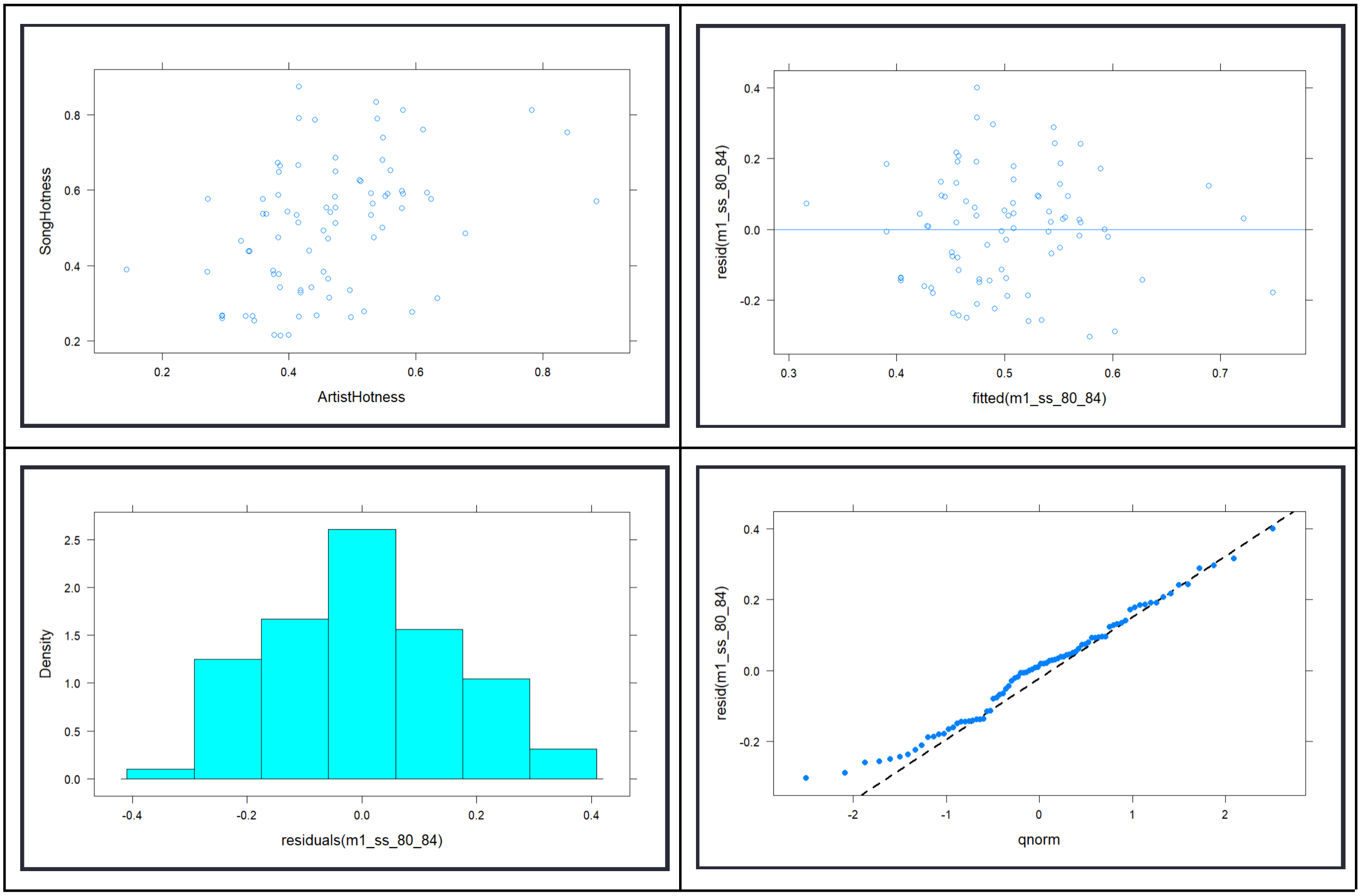
```
Call:
lm(formula = SongHotness ~ ArtistHotness, data = ss_85_89)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.35702 -0.11647 -0.00003  0.12721  0.30962 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)    0.25770    0.06607    3.90 0.000147 ***
ArtistHotness  0.56592    0.14220    3.98 0.000109 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1489 on 143 degrees of freedom
  (97 observations deleted due to missingness)
Multiple R-squared:  0.09971,	Adjusted R-squared:  0.09341 
F-statistic: 15.84 on 1 and 143 DF,  p-value: 0.0001093
```



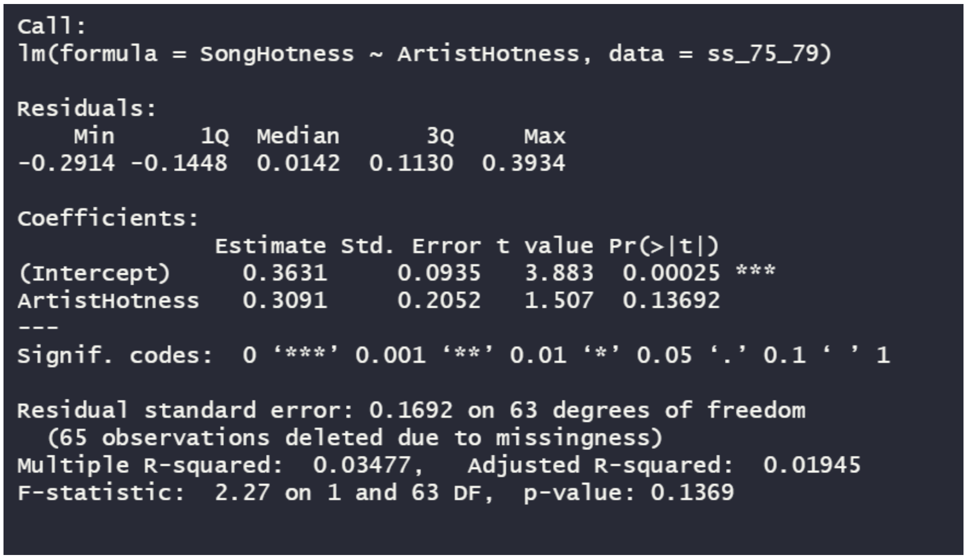
2. 1980 - 1984: We do not see a distinctly linear relationship between artist and song hotness. The R-squared value of 17% further indicates that a linear model is not suitable for modeling this relationship. The error plots reveal constant variance, zero mean, however the curved tail on the left in the normal probability plot indicates a deviation from the normal distribution.



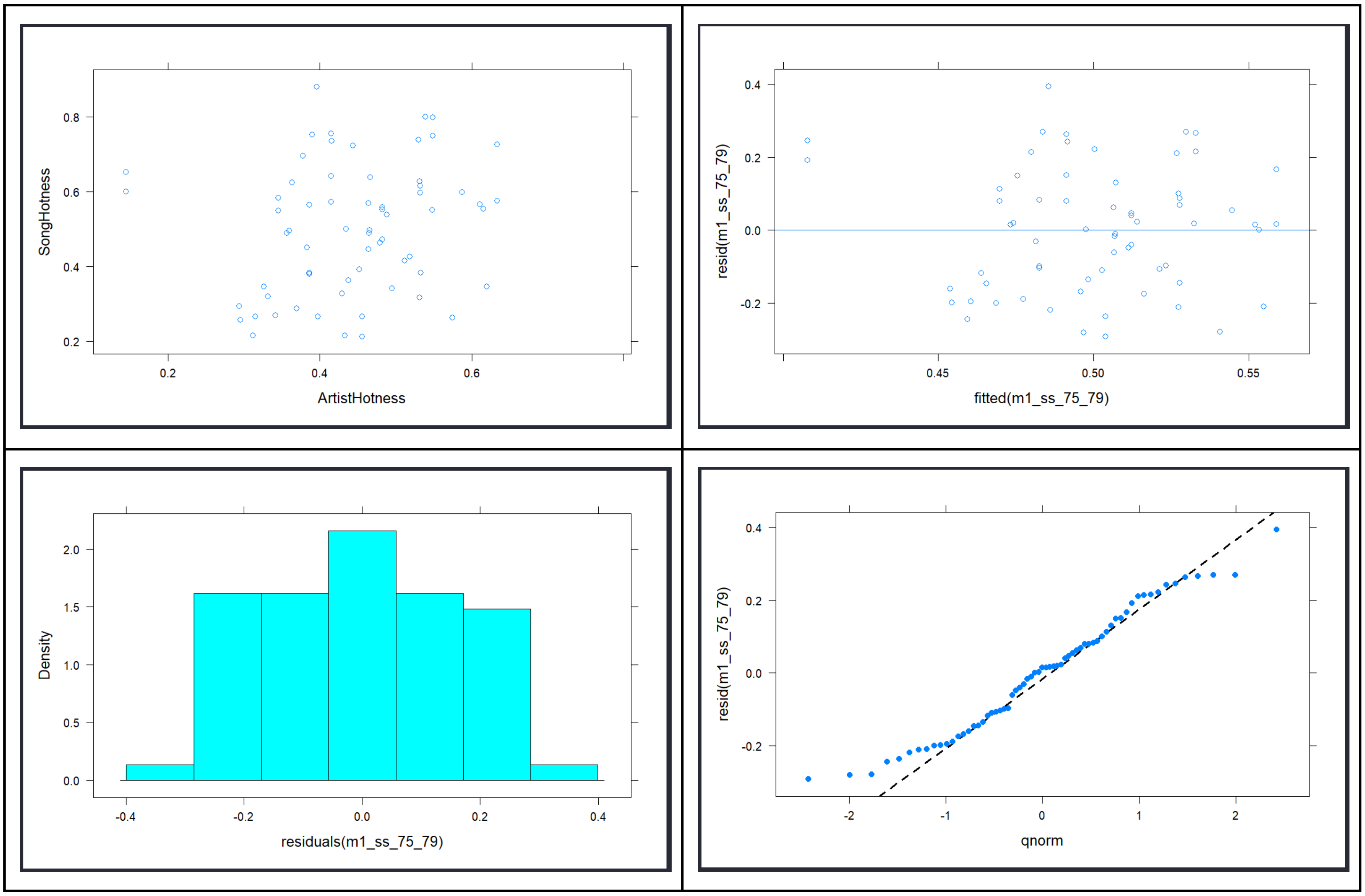
```
Call:
lm(formula = SongHotness ~ ArtistHotness, data = ss_80_84)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.30326 -0.13824  0.01358  0.09427  0.39904 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)    0.23227    0.06846   3.393 0.001078 ** 
ArtistHotness  0.58315    0.14295   4.080 0.000106 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1577 on 80 degrees of freedom
  (83 observations deleted due to missingness)
Multiple R-squared:  0.1722,	Adjusted R-squared:  0.1619 
F-statistic: 16.64 on 1 and 80 DF,  p-value: 0.000106
```



3. 1975 - 1979: We do not see a linear relationship between artist and song hotness. The R-squared value of 3% further indicates that a linear model is not suitable for modeling this relationship. The error plots reveal constant variance, zero mean, however the curved tails in the normal probability plot indicates a deviation from the normal distribution.



```
Call:
lm(formula = SongHotness ~ ArtistHotness, data = ss_75_79)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.2914 -0.1448  0.0142  0.1130  0.3934 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)     0.3631     0.0935   3.883  0.00025 ***
ArtistHotness   0.3091     0.2052   1.507  0.13692    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1692 on 63 degrees of freedom
  (65 observations deleted due to missingness)
Multiple R-squared:  0.03477,	Adjusted R-squared:  0.01945 
F-statistic:  2.27 on 1 and 63 DF,  p-value: 0.1369
```


SongHotness
ArtistHotness
resid(m1_ss_75_79)
fitted(m1_ss_75_79)
Density
residuals(m1_ss_75_79)
resid(m1_ss_75_79)
qnorm