The Billboard Hot 100 Songs Chart

ISYE 4031B
Group B5

Ndeyanta Jallow
Danielle Jones
Preston Akwule
Christian Hughey
Sunil Mutyala

Table of Contents

1. Introduction

Billboard has established itself to be a successful music media company within the United States. Individuals within this project group all have a shared interest pertaining to music that is consistently present on the "Billboard Hot 100." This is defined as the weekly record chart consisting of the top ranked 100 songs based on sales (physical and digital), radio play, and online streaming in the United States. A new chart is consistently compiled and released to the masses by *Billboard* every Tuesday. The rankings change rapidly throughout the year, thus, a diverse set of songs consistently populate this list. Many Data Scientists from companies such as Apple, Spotify, and Pandora have aimed to find the commonalities amongst successful songs primarily by doing Advanced Marketing Analytics that heavily relies upon regression and forecasting techniques.

The goal of this analysis is to focus primarily on the forecasting behind the "number of weeks" that respective songs have spent on the Billboard chart in the year of 2018. The analysis also aims to answer questions such as: What are the main factors that correlate to a longer "number of weeks" that a song has spent on the Billboard Hot 100? What are some commonalities between the songs that have been successful? How accurately can one predict the number of weeks that a song will stay on this list? Finally, it is important to note that the goal of this analysis is not to imply causation behind any trends; but rather, to establish the foundation for potential experiments behind music.

This analysis will incorporate both Simple Linear Regression and Multiple Linear Regression analysis techniques that will incorporate various screening techniques, residual analysis, data transformation, time series forecasting, qualitative data analysis, and smoothing methods to ultimately uncover the natural phenomena behind Billboard 100 song success.

2. Problem Statement

As avid music listeners, we are intrigued in music charts and how certain songs are consistently highlighted more than others. It was interesting as we began brainstorming ideas surrounding this topic, we discussed how some of our favorite songs were only listed on the Hot 100 list for one or two weeks where other songs that we didn't have much love for were on there for weeks on end. Upon realizing how much is put into music production, from the tempo of the song to the number of artists to the danceability, and the recognition that certain songs recieve, we became more and more interested in investigating the relationships between these independent variables and the number of weeks a song remains on the *Billboard*. We went further to identify exactly which independent variables showed statistical significance with respect to the dependent variable of the number of weeks a song remains on the *Billboard* Hot 100 chart.

As we dove into different sources of data (identified in the next section), we were able to compile the vast number of variables and songs into a more holistic compilation. The independent variables that this analysis focuses on are: *artist, peak_position, change, genre, tempo, duration_ms, feature, lifetime, number_of_artists, artist_origniation* and *single_vs_album*. We believe these variables are the most important to understanding the trends that currently exist in the music industry. As identified in the introduction, there are centralized questions that our analysis will delve into. However, our overarching goals are to identify the natural phenomena that occur in tandem with Billboard 100 music success while making structured recommendations behind how songs can remain on the chart longer without identifying causation.

There are certainly constraints that potentially exist with this analysis. The possible underrepresentation of the entire population of songs that exist out there in the music industry. We restricted our qualitative variable of "genre" to five different categories for the sake of simplicity. However, the music industry has evolved to the point where there are a multitude of genres that could be used to classify a song. Thus, insights based upon this specific could be potentially skewed.

The *artist* independent qualitative variable ended up being removed from this analysis because there were over 50 possible "values" (artist names) that appeared as corresponding options for the analysis. It was feasible to remove the variable *artist* who produced each song from the analysis, because the analysis included each artist's respective geographical region which is viewed to be a variable that is derived from the name of each arti

3. Data Description

The data compiled for this project was collected through both Spotify API's at https://developer.spotify.com/documentation/web-api/ as well as a spreadhseet on Data.world at https://data.world/typhon/Billboard-hot-100-songs-2000-2018-w-spotify-data-lyrics. As the original spreadsheet had over 7,500 rows and 31 columns, we knew for the scope of this project, we would not need to use all of this data. Therefore, we cleaned the spreadsheet to get a random sample of 150 songs that we used for the project. We first sorted the data to only be of the year 2018 as we wanted to specifically focus in on one recent and specific time frame. Then, we removed any rows that had "unknown" information so that it wasn't to be included in the final sample. After that, we decided upon which variables we wanted to include from the sheet in our model (*weeks, artist, peak_pos, change, tempo* and *duration_ms*). At this point, we had 206 data points remaining. In order to obtain a random sample of 150 out of the entire population, we wrote code in Python (shown in the appendix) to randomly select data points that we will include in our final spreadsheet of data. This was done to ensure that the data points used in the model were chosen at random, and not affected by any sort of bias. The other variables that we will have in our model (*feature, genre, lifetime, number_of_artists, artist_orginiation* and *single_vs_album*) were populated individually for each observation. We further inspected Spotify's web API to identify the information for the data points that were not originally prepopulated. Spotify is one of the top audio streaming platforms in the world with APIs that are used by many industry level software developers. Therefore, it is certainly viewed to be a credible source.

The sample size for the analysis is 150 observations. The spreadsheet here identifies each data point and its respective independent and dependent variables: https://docs.google.com/spreadsheets/d/1NIr-G3IovuvhGr7U6iqYNNQp4L9EInlndXrWmmqKKOg/edit?usp=sharing. In addition to the 150 observations used to build the model, we also have approximately 50 other observations that we plan to use to test the accuracy of it.

The following variables were used to predict the number of weeks in a year that a song will remain on the top 100 Billboard chart as the independent variables. They are a mixture of 8 quantitative and 2 qualitative variables. The qualitative variables were transformed to binary coding in our dataset. It is important to note that all observations under the *change* variable that were listed as "new" or "re-entering" could be defined to have a 0 associated with them. This is because songs that have not been on the Billboard 100 for consecutive weeks cannot exhibit a change in position. Thus, this was a significant data cleaning method.

Variable Number	Variable Name	Variable Type	Definition	Unit of Measure
1	*peak_position*	Quantitative	The highest number rank that the song reached in throughout the time it spent on the Billboard Hot 100 list.	Integer (positive only)
2	*change*	Quantitative	The net difference in the song's start position (the position of the song when it entered the Billboard Top 100) and song's end position (the song's last recorded position on the chart by the end of 2018 or the song's last position before leaving the Billboard Top 100).	Integer (positive or negative)
3	*genre*	Qualitative	The predetermined category of artistic composition that that song falls within	Possible values/categories: Country Hip Hop / Rap Misc Pop R&B
4	tempo	Quantitative	A value representing the speed of the song	Beats per minute
5	*duration*	Quantitative	The length of the song	Milliseconds
6	*feature*	Quantitative	Binary variable representing whether or not additional artists are featured on the track	0 - No Feature 1 - Feature

7	*lifetime*	Quantitative	How many years since the song was released from 2019 (ie: if a song was released in 2017, the lifetime would be 2)	Years
8	*number_of_artists*	Quantitative	The total number of individual artists in the song	Integer
9	*artist_origination*	Qualitative	The artist's place of birth	Possible Values: South, Northeast, Midwest, West, Pacific and International (The Geographical region corresponding to the artist's respective state)
10	*single_vs_album*	Quantitative	Binary variable indicating whether or not a song was represented on an album	0 - Single 1 - Album

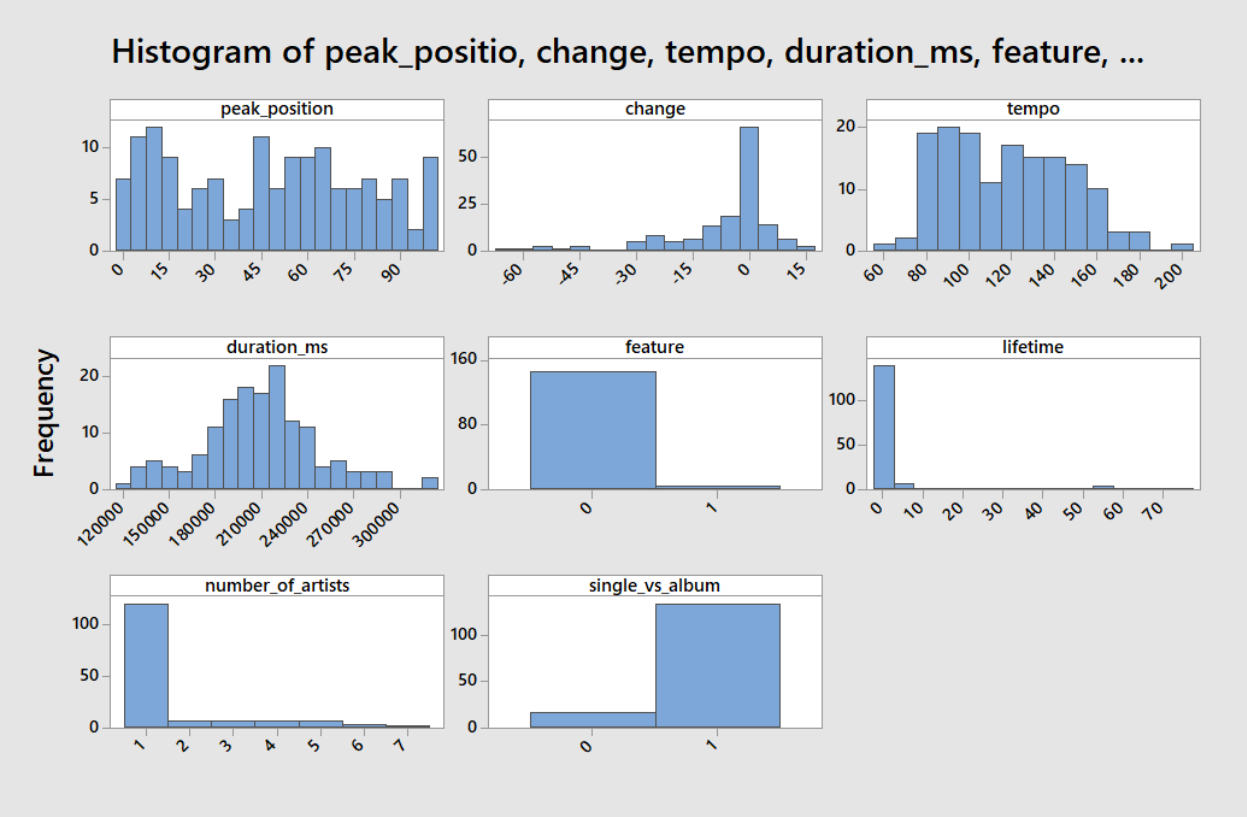
Dummy Variables: *Forecasting, Time Series and Regression* states that "we can model the effects of the different levels of a qualitative independent variable by using what we call dummy variables (also called indicator variables). Such variables are usually defined so that they take on two values—either 0 or 1." Thus, dummy variables were made to correspond to the both the genre and region qualitative variables. The base level for the genre dummy variable is R&B and the base level for the artist origination dummy variable is international.

4. Analysis

Screening Methods

The first step in this analysis involved creating a histogram of each quantitative variable to analyze their respective distributions for further exploration of the natural phenomena.

Histograms of Quantitative Variables



The above histograms of the quantitative variables all vary in their shape, symmetry, and spread, thus the interpretations of these varied from approximately normal distribution to heavily skewed. The variable *peak_position* was roughly uniform with several modes. Variable *change* was skewed left, unimodal, and seems to have outliers in the histogram. Variable *tempo* had a slight skew right. Variable *duration_ms* followed an approximately normal distribution and was unimodal. There were two binary variables we used in the model: variable *feature* was unimodal which indicates that a large majority of the data points reside with a value of "0" in the interval of the binary breakdown, and variable *single_vs_album* was unimodal which indicates that most of the data points have a value of "1" in the binary breakdown.

Variable *lifetime* was skewed right, unimodal, and appears to have extreme outliers. Lastly, variable *number_of_artists* was skewed right and unimodal.

The first step in this analysis was to investigate whether or not there was multicollinearity within independent variables. *Forecasting, Time Series and Regression* states that "Multicollinearity is said to exist among the independent variables in a regression situation if these independent variables are related to or dependent upon each other. One way to investigate multicollinearity is to examine the correlation matrix." Thus, a correlation matrix for this data set was produced as a result.

Correlation Coefficient Matrix

Correlation: weeks_on_chart, peak_position, change, ... mw, D_w, D_pa

Correlations

	weeks_on_chart	peak_position	change	tempo
peak_position	-0.531			
	0.000			
change	0.163	0.229		
	0.047	0.005		
tempo	-0.029	-0.124	-0.020	
	0.720	0.132	0.810	
duration_ms	0.001	0.021	-0.031	-0.218
	0.987	0.799	0.709	0.007
feature	-0.101	-0.070	-0.091	-0.010
	0.217	0.397	0.268	0.901
lifetime	0.061	-0.112	0.015	0.208
	0.455	0.172	0.855	0.011
number_of_artist	0.011	0.076	0.024	-0.111
	0.893	0.357	0.774	0.177
single_vs_album	0.164	-0.147	-0.055	0.115
	0.044	0.072	0.508	0.161
D_c	0.145	0.174	0.079	-0.246
	0.077	0.033	0.334	0.002
D_h	-0.325	0.018	-0.142	0.184
	0.000	0.831	0.083	0.024
D_m	0.079	0.068	0.094	-0.105
	0.336	0.411	0.251	0.200
D_p	0.191	-0.231	-0.004	0.059
	0.019	0.004	0.964	0.472
D_s	-0.171	0.271	-0.009	-0.023
	0.036	0.001	0.914	0.777
D_ne	-0.024	-0.140	-0.117	0.140
	0.770	0.089	0.154	0.087
D_mw	-0.023	0.052	0.034	0.185
	0.781	0.529	0.677	0.023
D_w	0.101	0.066	0.109	-0.113
	0.218	0.423	0.186	0.168
D_pa	0.284	-0.125	0.003	0.045
	0.000	0.129	0.970	0.589

Correlation Coefficient Matrix Continued

	duration_ms	feature	lifetime	number_of_artist
feature	-0.040			
	0.626			
lifetime	-0.215	-0.036		
	0.008	0.664		
number_of_artist	0.082	0.171	-0.082	
	0.319	0.036	0.321	
single_vs_album	-0.028	0.057	-0.104	0.055
	0.735	0.487	0.206	0.501
D_c	-0.025	-0.083	-0.062	-0.120
	0.759	0.314	0.453	0.144
D_h	-0.234	0.224	-0.154	-0.138
	0.004	0.006	0.060	0.093
D_m	-0.061	-0.051	0.384	0.301
	0.455	0.535	0.000	0.000
D_p	0.277	-0.117	-0.010	0.087
	0.001	0.154	0.906	0.292
D_s	-0.173	0.020	0.019	-0.061
	0.034	0.808	0.816	0.460
D_ne	0.179	0.018	-0.058	-0.162
	0.028	0.830	0.483	0.048
D_mw	-0.120	-0.047	0.095	-0.047
	0.145	0.572	0.247	0.564
D_w	0.024	-0.063	0.046	0.364
	0.767	0.443	0.576	0.000
D_pa	-0.004	-0.014	-0.001	-0.036
	0.960	0.869	0.992	0.665

Correlation Coefficient Matrix Continued

	single_vs_album	D_c	D_h	D_m
D_c	0.011			
	0.896			
D_h	-0.106	-0.370		
	0.197	0.000		
D_m	-0.124	-0.154	-0.228	
	0.131	0.060	0.005	
D_p	0.153	-0.354	-0.523	-0.218
	0.062	0.000	0.000	0.007
D_s	-0.042	0.396	0.131	-0.130
	0.612	0.000	0.109	0.113
D_ne	0.123	-0.255	-0.102	-0.040
	0.134	0.002	0.215	0.625
D_mw	-0.068	-0.077	0.060	0.004
	0.405	0.351	0.469	0.959
D_w	-0.063	-0.090	-0.072	0.310
	0.442	0.272	0.382	0.000
D_pa	0.028	-0.041	-0.061	-0.025
	0.731	0.619	0.462	0.759

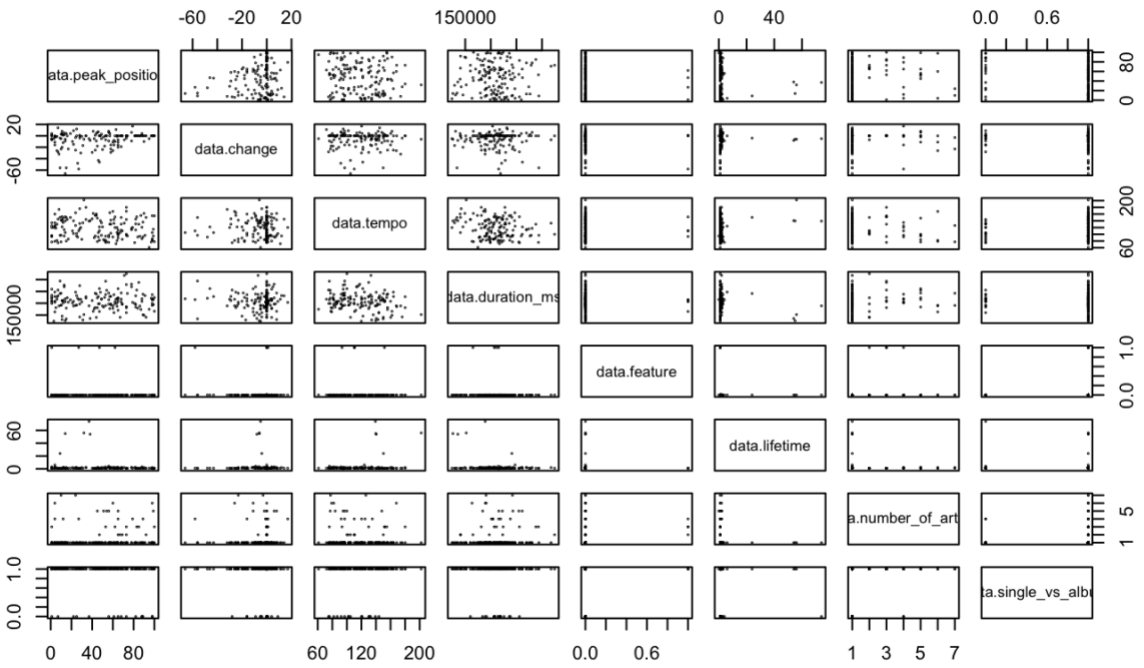
	D_p	D_s	D_ne	D_mw
D_s	-0.342			
	0.000			
D_ne	0.338	-0.452		
	0.000	0.000		
D_mw	-0.090	-0.249	-0.144	
	0.271	0.002	0.080	
D_w	-0.014	-0.338	-0.194	-0.107
	0.863	0.000	0.017	0.192
D_pa	0.116	-0.073	-0.042	-0.023
	0.158	0.377	0.611	0.780

	D_w
D_pa	-0.031
	0.705

We analyzed the correlation matrix above with the purpose of identifying variables that are highly correlated in the current model. We chose an alpha value of 0.01 to identify which correlation coefficients are significantly different from 0 (if $p<0.01$) and identified 28 pairs of significant variables in matrix above. In examining the first column which shows relationships to the response variable, only one of the correlation coefficients has a moderately high value(above |0.5|: [*peak_position, weeks_on_chart*] with a correlation coefficient of -0.531. This shows that there could be a significant correlation between the response variable and the *peak_position* independent variable. In examining the correlation matrix outside of the first column, the correlation coefficients between the independent variables, we can identify possible relationships that exhibit multicollinearity. A few notable relationships that have coefficients significantly different from 0 are [D_h, D_p] and [D_s, D_ne] which both have values above |0.4| but less than |0.6|. Additionally,

Forecasting, Time Series and Regression asserts that **"**statisticians often regard multicollinearity in a data set to be severe if at least one simple correlation coefficient between the independent variables is at least 0.9." Therefore, the relationships between variables exhibited no multicollinearity within this data set. There were no correlation coefficients above 0.9. We will further explore how the correlation between these variables are affected as we build new models and remove influential points.

Correlation Plots



All Variables Model

To get a holistic starting point to see where future iterations could be improved, the group analyzed the model with all of the variables initially included. (See appendix for results of multilinear regression executed.)

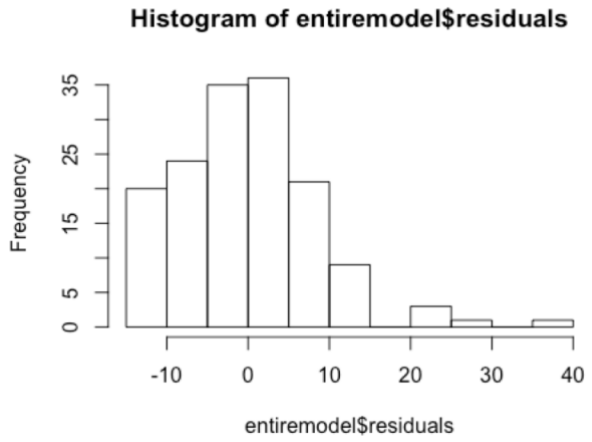
Output from All Variables

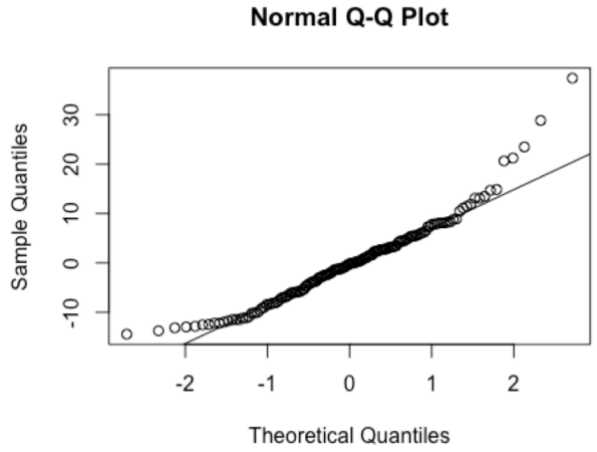
```
Call:
lm(formula = y ~ position + duration + feature + lifetime + numartists +
    singleoralbum + country + hiprap + misc + pop + south + northeast +
    midwest + west + pacific + tempo)

Residuals:
    Min      1Q  Median      3Q     Max
-14.467  -6.002  -0.243   4.517  37.404

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.552e+01  8.685e+00   1.787  0.07618 .
position      -2.140e-01  2.778e-02  -7.702 2.72e-12 ***
duration      -9.205e-06  2.193e-05  -0.420  0.67530
feature       -4.361e+00  4.934e+00  -0.884  0.37835
lifetime      -1.007e-01  9.341e-02  -1.078  0.28300
numartists    -3.198e-01  6.649e-01  -0.481  0.63133
singleoralbum  3.196e+00  2.534e+00   1.261  0.20943
country        1.057e+01  5.233e+00   2.020  0.04544 *
hiprap         1.250e-02  4.989e+00   0.003  0.99801
misc           9.885e+00  5.828e+00   1.696  0.09217 .
pop            5.851e+00  4.991e+00   1.172  0.24315
south         -2.532e-01  2.588e+00  -0.098  0.92222
northeast     -1.653e+00  2.658e+00  -0.622  0.53503
midwest        2.969e+00  3.658e+00   0.812  0.41852
west           4.217e+00  3.138e+00   1.344  0.18139
pacific        3.029e+01  9.339e+00   3.243  0.00149 **
tempo         -8.994e-04  2.948e-02  -0.031  0.97571
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.036 on 133 degrees of freedom
Multiple R-squared:  0.4893,    Adjusted R-squared:  0.4278
F-statistic: 7.963 on 16 and 133 DF,  p-value: 5.059e-13
```

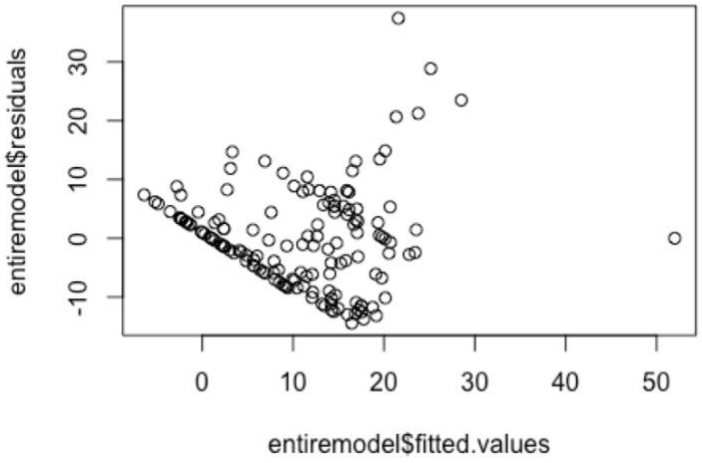
ANOVA from All Variables Model

```
Analysis of Variance Table

Response: y
               Df  Sum Sq Mean Sq F value    Pr(>F)
position        1  6000.7  6000.7 73.4974  2.25e-14 ***
duration        1     3.3     3.3  0.0407 0.8404084
feature         1   406.5   406.5  4.9785 0.0273393 *
lifetime        1     0.2     0.2  0.0021 0.9637424
numartists      1   125.6   125.6  1.5389 0.2169721
singleoralbum   1   172.7   172.7  2.1153 0.1481864
country         1  1228.8  1228.8 15.0504 0.0001639 ***
hiprap          1   853.3   853.3 10.4513 0.0015456 **
misc            1   201.8   201.8  2.4717 0.1182900
pop             1    84.0    84.0  1.0291 0.3122177
south           1    64.1    64.1  0.7850 0.3772060
northeast       1   310.5   310.5  3.8026 0.0532764 .
midwest         1     1.8     1.8  0.0219 0.8825617
west            1    87.3    87.3  1.0695 0.3029428
pacific         1   861.8   861.8 10.5561 0.0014665 **
tempo           1     0.1     0.1  0.0009 0.9757060
Residuals     133 10858.7    81.6
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Residual Analysis Plots for Entire Model (Histogram, QQPlot, Scatterplot)







Through the analysis of the initial model summary and residuals it was apparent that more iterations would be necessary to reach a better model. First, there was only one variable that was shown to be significant based upon the column titled "Pr(>| t |)" represents the p-values for each respective model variable. When determining if a variable is statistically significant or not, we looked at the magnitude of the corresponding p-value and compared it to the appropriate significance level. In our model, we are testing at a significance level of 0.05, therefore any p-value for a model variable less than the significance level is recognized as statistically significant. Thus, only the position variable proved to be helpful.

The residual analysis also shows that this model could be improved. "To check the validity of the constant variance assumption, we examine plots of the residuals against the fitted values. When we look at these plots, the pattern of the residuals' has a strong linear decreasing trend from 10 down to -10 violates the constant variance assumption and the independence assumption because of the strong pattern exhibited. A residual plot that "fans out" suggests that the error terms are becoming more spread out as the horizontal plot value increases and that the constant variance assumption is violated. Here we would say that an increasing error variance exists." This was shown to be the case when looking at the residuals of the model including all of the variables. The histogram above shows that the residuals also violate the normality assumption because it was skewed to the right. The mean of the residuals was shown to be -7.50511 x 10^-16. There is no autocorrelation present either. The Durbin-Watson test produced a value of 2.013 which is slightly greater than 2, this hints at a slight presence of negative autocorrelation.

The model outputted from R is as follows: $\hat{y}$ = **1.552e01 - 2.14e-01**position **- 9.205e-06**duration **- 4.361e00**feature **- 1.007e-01**lifetime **- 3.198e-01**numartists + **3.196e00**singleoralbum + **1.057e01**country + **1.250e-02**hiprap + **9.885e00**misc + **5.851e00**pop **-2.532e-01**south **-1.653e00**northeast +**2.969e00**midwest + **4.217e00**west +**3.029e01**pacific **-8.994e04**tempo

Stepwise Regression

In order to build a model that was truly statistically significant and representative of our data, we ran three different methods that are commonly used to develop models. Using R, we first manipulated our data to include the dummy variables explained above for both genre and artist origination. Then, through the use of the respective regression packages, we were able to employ the iterative model selection procedures - detailed further in the appendix. The initial model with all the variables is outlined below:

Initial Model:
y ~ position + duration + feature + lifetime + numartists + singleoralbum + country + hiprap + misc + pop + south + northeast + midwest + west + pacific

Upon the use of the stepwise regression method in R, the output model produced was:
y ~ position + country + misc + pop + west + pacific

Since the *country* variable was viewed as significant, each of the following were as well: *hiprap*, *misc* and *pop*. Likewise, since both the *west* and *pacific* variables were viewed as significant, all of the categories within the artist origination were also significant: *south*, *northeast* and *midwest*. Both of these occurrences with the two different variables was due to the principle of hierarchy.

Thus, the first model from this iterative method based upon stepwise regression would be a 10 variable model: **y** ~ position + country + hiprap + misc + pop + south + northeast + midwest + west + pacific

Output from Stepwise Regression

```
Call:
lm(formula = y ~ position + country + hiprap + misc + pop + south +
    northeast + midwest + west + pacific)

Residuals:
    Min      1Q  Median      3Q     Max
-15.194  -5.668  -0.119   4.605  37.889

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 16.17882    5.07950   3.185  0.00179 **
position    -0.21291    0.02665  -7.989 4.66e-13 ***
country     10.35159    5.12907   2.018  0.04549 *
hiprap      -0.64169    4.88568  -0.131  0.89570
misc         7.34670    5.40161   1.360  0.17600
pop          5.14472    4.90583   1.049  0.29614
south       -0.48139    2.47353  -0.195  0.84598
northeast   -1.33870    2.54849  -0.525  0.60022
midwest      2.57441    3.46341   0.743  0.45854
west         4.01951    2.99522   1.342  0.18179
pacific     30.88938    9.23701   3.344  0.00106 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.976 on 139 degrees of freedom
Multiple R-squared:  0.4732,    Adjusted R-squared:  0.4353
F-statistic: 12.49 on 10 and 139 DF,  p-value: 2.691e-15
```

To confirm the validity of the best subsets regression method the group analyzed the model summary of the corresponding output. The far most right column represents the p-value of each coefficient corresponding to the model variables. A p-value less than 0.05 (standard alpha value) corresponds to a significant predictor. Based upon this criteria the *country, position,* and *pacific* variables were shown to be significant.

Backwards Elimination

The backwards regression method produced a collinear output which supports the results of the stepwise regression method.

Initial Model:

y ~ position + duration + feature + lifetime + numartists + singleoralbum + country + hiprap + misc + pop + south + northeast + midwest + west + pacific

Output Model:

y ~ position + country + misc + pop + west + pacific

Thus, applying the same logic behind the principle of hierarchy, one can see that the same 10 variable model is produced.

y ~ position + country + hiprap + misc + pop + south + northeast + midwest + west + pacific

Output from Backwards Elimination

```
Call:
lm(formula = y ~ position + country + hiprap + misc + pop + south +
    northeast + midwest + west + pacific)

Residuals:
    Min      1Q  Median      3Q     Max
-15.194  -5.668  -0.119   4.605  37.889

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 16.17882    5.07950   3.185  0.00179 **
position    -0.21291    0.02665  -7.989 4.66e-13 ***
country     10.35159    5.12907   2.018  0.04549 *
hiprap      -0.64169    4.88568  -0.131  0.89570
misc         7.34670    5.40161   1.360  0.17600
pop          5.14472    4.90583   1.049  0.29614
south       -0.48139    2.47353  -0.195  0.84598
northeast   -1.33870    2.54849  -0.525  0.60022
midwest      2.57441    3.46341   0.743  0.45854
west         4.01951    2.99522   1.342  0.18179
pacific     30.88938    9.23701   3.344  0.00106 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.976 on 139 degrees of freedom
Multiple R-squared:  0.4732,    Adjusted R-squared:  0.4353
F-statistic: 12.49 on 10 and 139 DF,  p-value: 2.691e-15
```

The column titled "Pr(>| t |)" represents the p-values for each respective model variable. When determining if a variable was statistically significant or not, we looked at the magnitude of the corresponding p-value and compared it to the appropriate significance level. In our model, we were testing at a significance level of 0.05, therefore any p-value for a model variable less than the significance level was recognized as statistically significant; however, variables with p-values greater than the significance level were not significant. According to this criterion, variables *position*, *country*, and *pacific* were statically significant variables to our model. As you can see, the statistically significant variables for the stepwise and backwards regression were consistent in both regression methods.

Best Subsets

With best subsets, we chose to represent the first two best models with 10 variables to remain consistent with the results in the number of variables from stepwise and backwards regression.

First best model: y ~ position + feature + lifetime + singleoralbum + country + hiprap + misc + pop + south + northeast + midwest + west + pacific

Second best model: y ~ position + feature + singleoralbum + country + hiprap + misc + pop + south + northeast + midwest + west + pacific

Output from Best Subsets(First Model)

```
Call:
lm(formula = y ~ position + feature + lifetime + singleoralbum +
    country + hiprap + misc + pop + south + northeast + midwest +
    west + pacific)

Residuals:
    Min      1Q  Median      3Q     Max
-15.585  -5.968  -0.376   4.362  37.561

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  13.03329    5.73568   2.272  0.02464 *
position     -0.21519    0.02735  -7.867 1.01e-12 ***
feature      -5.01447    4.69891  -1.067  0.28779
lifetime     -0.08263    0.08553  -0.966  0.33568
singleoralbum  3.20247    2.46573   1.299  0.19621
country      10.73071    5.12517   2.094  0.03814 *
hiprap        0.25364    4.90777   0.052  0.95886
misc          9.41868    5.56085   1.694  0.09260 .
pop           5.58377    4.90321   1.139  0.25679
south        -0.28163    2.50230  -0.113  0.91055
northeast    -1.52536    2.54878  -0.598  0.55052
midwest       3.03534    3.50957   0.865  0.38863
west          3.87064    2.99238   1.293  0.19803
pacific      30.64356    9.21081   3.327  0.00113 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.95 on 136 degrees of freedom
Multiple R-squared:  0.4876,    Adjusted R-squared:  0.4387
F-statistic: 9.956 on 13 and 136 DF,  p-value: 2.014e-14
```

This model produced an output that was in agreement with the stepwise and backwards regression methods. This was not the initial case. The output initially produced a model consisting of (y ~ position + feature + lifetime + singleoralbum + country + hiprap + misc + pop + south + northeast + midwest + west + pacific). However, after looking at the model summary the lifetime, feature, and singleoralbum variables were taken out because they had p-values that were greater than 0.05. Thus, when the variables were removed the variables chosen were the same as the ones aforementioned in the backwards and stepwise regression sections.

"Best choice" out of screening methods

After running screening methods on all the variables we then ran multiple linear regression models on each output. The best subsets model, backwards elimination and stepwise models concluded that three independent variables were significant in determining the number of weeks a song remained on the *Billboard* (position, region, and artist origination). We then tested the hypothesis:

H_0: $B_j = 0$ $\quad j = 1,2,..,10$

H_a: $B_j \neq 0$

After conducting p-value tests against an alpha of 0.05 we rejected the null and concluded that at least one of the variables was significant in predicting the number of weeks a song would remain on a chart. The independent variables deemed significant were: position, genre, and artist origination by location. The best subsets model in R studio produced the highest adjusted R^2 value (43.87%) and the lowest residual standard error of 8.95. From this model we obtained the resulting regression equation:

$\hat{y}$ = **13.02** - **0.22**position + **10.73**country - **0.25**hip-hop/rap + **9.42**misc + **5.58**pop - **0.28**south - **1.53**northeast + **3.04**midwest + **3.87**west + **30.64**pacific

Since the R^2 value for this regression model was 48.76%, we concluded that 48.76% of the total model variation can be explained by the linear relationship between the number of weeks on the chart and the significant independent variables. We observed that this percentage was low and so we ran further iterations and diagnostics to try and improve the model and fix the issues. After re-running the model in Minitab, we compared the Mallow's C_p values of each subset. We observed that upon including the independent variable "change" in the model, the adjusted R^2 value increased to 49.6% (R^2 of 51.7%) with the residual standard error (s) reducing to 8.48. This model concluded that 6 variables were useful as per the $C_p < p+1$ criterion ($4.9 < 6+1$):

Output from Best Subsets Including Change Variable

Best Subsets Regression: weeks_on_chart versus ... e, D_mw, D_w, D_pa

Response is weeks_on_chart

Vars	R-Sq	R-Sq (adj)	R-Sq (pred)	Mallows Cp	S	peak_position	change	D_c	D_h	D_m	D_p	D_s	D_ne	D_mw	D_w	D_pa
1	28.2	27.7	25.9	63.3	10.154	X										
1	10.6	10.0	8.5	114.7	11.334				X							
2	38.2	37.4	35.5	36.1	9.4534	X			X							
2	36.8	35.9	34.2	40.4	9.5633	X	X									
3	44.3	43.2	41.4	20.3	9.0029	X	X		X							
3	42.2	41.1	*	26.4	9.1707	X			X							X
4	48.1	46.7	*	11.2	8.7199	X	X		X							X
4	46.6	45.2	*	15.6	8.8464	X	X	X								X
5	50.2	48.5	*	7.1	8.5722	X	X	X	X							X
5	49.6	47.9	*	8.8	8.6227	X	X		X		X					X
6	51.7	49.6	*	4.9	8.4763	X	X	X	X						X	X
6	51.3	49.3	*	5.8	8.5049	X	X	X		X	X					X
7	52.3	49.9	*	5.1	8.4516	X	X	X		X	X				X	X
7	52.0	49.6	*	6.1	8.4817	X	X	X	X	X					X	X
8	52.6	49.9	*	6.1	8.4509	X	X	X		X	X			X	X	X
8	52.3	49.6	*	6.9	8.4771	X	X	X		X	X		X		X	X
9	52.7	49.6	*	8.0	8.4792	X	X	X		X	X	X		X	X	X
9	52.7	49.6	*	8.0	8.4797	X	X	X		X	X		X	X	X	X
10	52.7	49.3	*	10.0	8.5089	X	X	X	X	X	X	X		X	X	X
10	52.7	49.3	*	10.0	8.5092	X	X	X	X	X	X		X	X	X	X
11	52.7	48.9	*	12.0	8.5395	X	X	X	X	X	X	X	X	X	X	X

At this point we then concluded that the significant independent variables in determining the number of weeks on the chart were: peak position, change, genre and artist origination (location). We then ran a multiple linear regression model with the change variable included and after comparing the p-value of the change variable we concluded that it was indeed significant with regards to the whole model. Testing the significance of the regression relationships between y and x values we also used the overall F statistic test, F(model). We tested the aforementioned hypothesis and defined the rejection region as:

If: $F(\text{model}) > F_{[a]}$, we reject H_0 in favor of H_a

$F(\text{model}) \leq F_{[a]}$, we fail to reject H_0

$$F(\text{model}) = \frac{(Explained\ variation)/k}{(Unexplained\ variation)/[n-(k+1)]}$$

$$= \frac{1017.98}{72.92}$$

$$= 13.96$$

$F_{[a]} \sim 1.8721$

Since F(model) > $F_{[a]}$ it is safe to say we can reject H_0 in favor of H_a and conclude that the model is significant. This is intuitive because obtaining a large overall F-statistic should be obtained when comparing the ratio of explained variation to unexplained variation. After making these adjustments based on our conclusions we obtained the following regression equation and model summary:

Regression Equation and Model Summary

Regression Equation

weeks_on_chart = 17.21 - 0.2362 peak_position + 0.2065 change + 10.89 D_c + 0.72 D_h + 7.63 D_m + 5.72 D_p + 0.29 D_s - 0.20 D_ne + 2.96 D_mw + 4.04 D_w + 30.54 D_pa

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
8.53951	52.67%	48.89%	*

Unusual and Influential Points

Unusual and Influential Points

Fits and Diagnostics for Unusual Observations

Obs	weeks_on_chart	Fit	Resid	Std Resid		
12	42.00	21.11	20.89	2.40	R	
19	59.00	21.11	37.89	4.35	R	
42	11.00	15.33	-4.33	-0.58		X
67	52.00	26.69	25.31	3.01	R	
72	1.00	2.70	-1.70	-0.23		X
73	54.00	25.68	28.32	3.32	R	
88	4.00	1.29	2.71	0.36		X
142	1.00	-2.33	3.33	0.44		X
144	52.00	52.00	0.00	*		X
146	45.00	22.19	22.81	2.73	R	

R Large residual
X Unusual X

Obs	Cook's D	DFITS		
12	0.03	0.61091	R	
19	0.11	1.16729	R	
42	0.01	-0.39662		X
67	0.11	1.14925	R	
72	0.00	-0.14696		X
73	0.11	1.12504	R	
88	0.00	0.22204		X
142	0.01	0.27873		X
144	*	*		X
146	0.10	1.08771	R	

R Large residual
X Unusual X

Durbin-Watson Statistic

Durbin-Watson Statistic = 2.02404

A |SRES| > 2 is a sign of an observation may be an outlier. Thus, the 12th, 19th, 67th, and 146th observations were taken out based upon this criteria because they had values of 2.40, 4.35, 3.32, and 2.76 respectively. The Minitab output also recommended that all of the other points that were "unusual" should be taken out.

Any Cook's Distance values of $D_i > 1$ or $D_i > 4/n$ or $D_i > 4/(n-(k+1)$ was suspected to be influential (*where k=4 and n=150*). Therefore, this information was taken into account when making changes to the model. The choice to remove these unusual observations and influential points was validated through the output. More of the variation in Billboard 100 chart success was explained through the removal of influential points and unusual observations.

Output from Removing Unusual/Influential Points with Change Variable

```
Call:
lm(formula = newy ~ newposition + newcountry + newhiprap + newmisc +
    newpop + newsouth + newnortheast + newmidwest + newwest +
    newpacific + newchange)

Residuals:
    Min      1Q  Median      3Q     Max
-14.575  -4.002  -1.078   3.626  15.920

Coefficients: (2 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  16.67482    1.83533   9.085 1.47e-15 ***
newposition  -0.18247    0.02141  -8.521 3.42e-14 ***
newcountry    4.65725    1.82173   2.557  0.01172 *
newhiprap    -4.12035    1.43170  -2.878  0.00468 **
newmisc      -2.30305    2.33915  -0.985  0.32666
newpop             NA         NA      NA       NA
newsouth      2.84051    1.93021   1.472  0.14354
newnortheast  3.78417    1.99816   1.894  0.06047 .
newmidwest    5.85300    2.73235   2.142  0.03405 *
newwest       6.88736    2.29721   2.998  0.00326 **
newpacific         NA         NA      NA       NA
newchange     0.17099    0.04079   4.192 5.07e-05 ***
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.608 on 130 degrees of freedom
Multiple R-squared:  0.4824,    Adjusted R-squared:  0.4465
F-statistic: 13.46 on 9 and 130 DF,  p-value: 4.555e-15
```

Assessment of Interaction Variables

Influential points and outliers were observed after running the Cook's D, TRES and SRES tests in the section above. We then went further to test for interaction variables based on the correlation matrix produced after removing these unusual points. After obtaining this model we realized it was dangerous to ignore the possibility of significant interaction terms. We then went through and added some cross products of quantitative variables to this regression model one at a time to test for significance. Among the cross products tried were:

- peak_position*change
- peak_position*duration
- peak_position*tempo
- peak_position*lifetime
- change*duration
- change*tempo
- change*lifetime

These new variables were tested for significance because peak_position and change were the two quantitative variables that we already concluded had a significant effect on the regression model. As such, we ran 7 more regression models respectively. The p-value tests on the additional variables, concluded that none of these variables contributed to the overall improvement of the model. It was safe to say that we could move on with the model obtained above to check for transformations.

Regression Equation

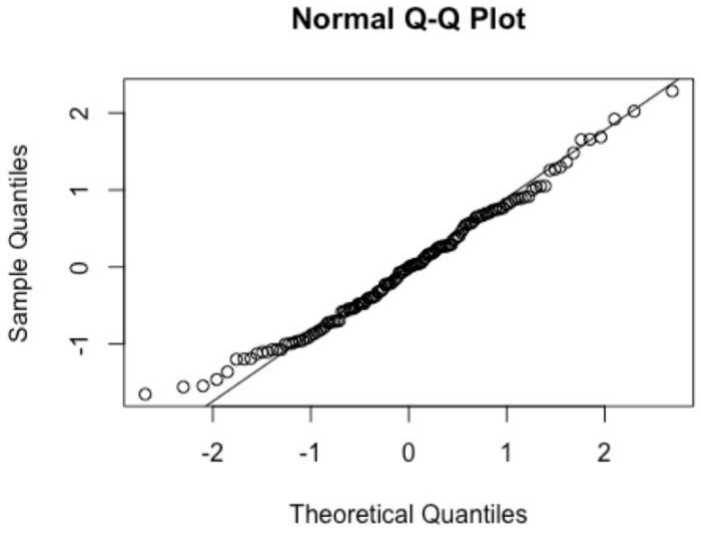
Regression Equation

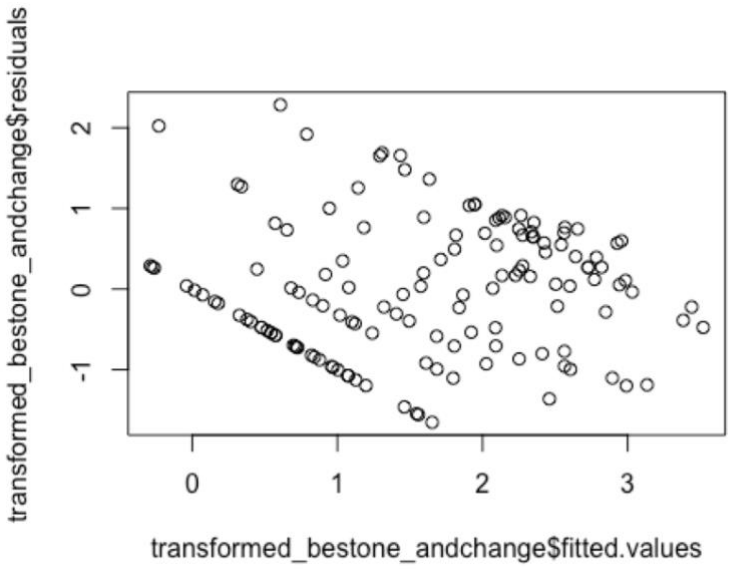
weeks_on_chart = 17.21 - 0.2362 peak_position + 0.2065 change + 10.89 D_c + 0.72 D_h + 7.63 D_m + 5.72 D_p + 0.29 D_s - 0.20 D_ne + 2.96 D_mw + 4.04 D_w + 30.54 D_pa

At the end of this analysis, we stuck to our findings that out of all the variables tested, the independent variables: peak_position, change, location and genre were significant in predicting the number of weeks on the chart.

Natural Log Transformation of Regression Model

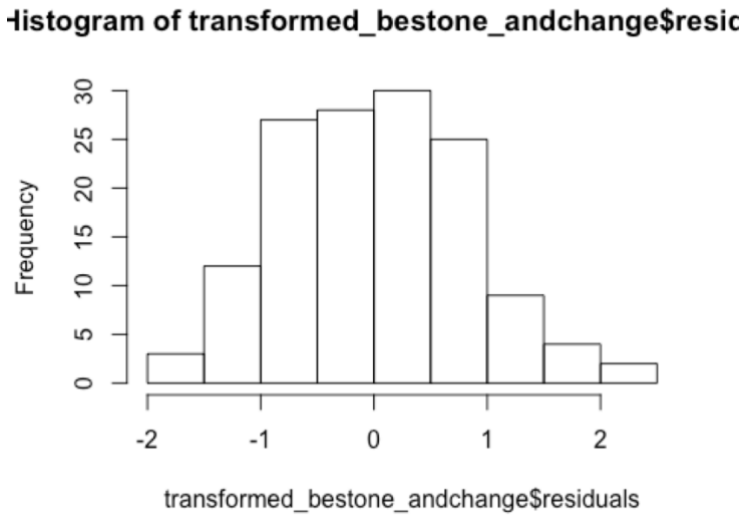
QQ-Plot, Model Output, ANOVA, Histogram and Residual Plot for Transformed data





```
Call:
lm(formula = log(newy) ~ newposition + newcountry + newhiprap +
    newmisc + newpop + newsouth + newnortheast + newmidwest +
    newwest + newpacific + newchange)

Residuals:
     Min       1Q   Median       3Q      Max
-1.65303 -0.57841 -0.00374  0.60934  2.28346

Coefficients: (2 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.744816   0.234574  11.701  < 2e-16 ***
newposition -0.027745   0.002737 -10.138  < 2e-16 ***
newcountry   0.763459   0.232835   3.279  0.00134 **
newhiprap   -0.598959   0.182986  -3.273  0.00136 **
newmisc     -0.103654   0.298967  -0.347  0.72937
newpop            NA         NA      NA       NA
newsouth     0.284029   0.246700   1.151  0.25172
newnortheast 0.354882   0.255385   1.390  0.16703
newmidwest   0.950438   0.349222   2.722  0.00739 **
newwest      0.955031   0.293607   3.253  0.00146 **
newpacific        NA         NA      NA       NA
[illegible]
newchange    0.016317   0.005213   3.130  0.00216 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8445 on 130 degrees of freedom
Multiple R-squared:  0.5556,    Adjusted R-squared:  0.5249
F-statistic: 18.06 on 9 and 130 DF,  p-value: < 2.2e-16
```

```
Analysis of Variance Table

Response: log(newy)
             Df Sum Sq Mean Sq F value    Pr(>F)
newposition   1 60.554  60.554 84.9047 7.099e-16 ***
newcountry    1 22.793  22.793 31.9589 9.529e-08 ***
newhiprap     1 12.963  12.963 18.1757 3.842e-05 ***
newmisc       1  0.283   0.283  0.3964  0.530072
newsouth      1  1.800   1.800  2.5239  0.114564
newnortheast  1  1.132   1.132  1.5870  0.210017
newmidwest    1  1.591   1.591  2.2313  0.137665
newwest       1  7.823   7.823 10.9682  0.001200 **
newchange     1  6.988   6.988  9.7976  0.002158 **
Residuals   130 92.716   0.713
---
```



The histogram of the of the transformed data shown above is approximately normal in distribution, unimodal, and fairly symmetrical in shape and spread. In the QQ plot we can see that majority of the data points lie on or near a straight line with a conglomeration of data points in the center of the line. There appears to be some spread as the data points go further along the ends of the line. Notice that the x-axis plots are the theoretical quantiles and those are from the Standard Normal distribution with mean 0 and standard deviation 1. Overall it would be fair to assume the distribution is fairly normal and consistent with the histogram.

The p-values for each respective model variable are given in the last column to the right. When comparing the magnitude of the corresponding p-values of their respective model variable to the significance level of 0.05, we see that variables *newposition, newcountry, newhiprap, newmidwest, newwest and newchange* have p-values less than 0.05 are deemed statistically significant to the model. As per the other models, the entire genre and artisitorginiation variables become significant since at least one of the categories within each variable are significant (ie: newhiprap and newmidwest). The rest of the variables with p-values greater than the significance level would be not considered statistically significant.

The final model produced adjusted and multiple R squared values of 52.49% and 55.56% respectively. These values show that through multiple iterations of changes and transformations an improved model was ultimately achieved.

It is important to note that the final model had "NA" in the output for dummy variables because the only observations with "1" as the value were taken out as unusual observations/influential points.

The following model has been developed for a transformed y:

$\hat{y}^* = 2.70 - 0.27$newposition $+ 0.76$newcountry $- 0.59$newhip-hop/rap $- 0.10$newmisc $+ 0$newpop $+ 0.28$newsouth $+ 0.35$newnortheast $+ 0.95$newmidwest $+ 0.95$newwest $+ 0$newpacific $+ 0.016317$newchange

To be able to compare this to the original model, we have to solve for $\hat{y}$ by taking inverse of the natural log of the entire model.

The model becomes:

$\hat{y} = 14.88 - 1.31$newposition $+ 2.13$newcountry $- 1.80$newhip-hop/rap $- 1.11$newmisc $+ 0$newpop $+ 1.32$newsouth $+ 1.42$newnortheast $+ 2.59$newmidwest $+ 2.59$newwest $+ 0$newpacific $+ 1.02$newchange

5. Conclusion

Model quality was ultimately improved between the beginning of the iteration process to the end. The group started off by analyzing a model that encompassed the 10 variables included in the data description section. This model did not meet the quality that the group aimed to seek when it came to predicting the number of weeks that a Billboard Hot 100 song will spend on the chart throughout the year. The main indicators of the lack of quality within the initial model were exposed through the breaking of the constant variance assumption for the residuals (shown through the fan pattern on the residual plot of the initial model) and the low adjusted and multiple R-squared values of 43.21% and 48.93% respectively. The standard error of the initial model was 9.0332 as well. Through multiple iterations of analysis the group achieved a better model that still violated the error of the residuals. However, the violation in the final model is much less significant than that of the initial model. The final model also has higher adjusted and multiple R squared values of 52.49% and 55.56% respectively. The standard error of the final model was 0.8444 as well. Thus, a model that was transformed with a Natural Log function containing only significant predictors was achieved and it was much better in quality than the initial model when looking at the aforementioned comparison metrics. Thus, this model explains more variation in Billboard Hot 100 chart success than that of the initial one.

Beginning our comparison of the correlation analysis of the variables in our initial model versus our final model, we looked at the first column which shows the possible strength of relationships between independent variables and the response variable. As with the initial model, the highest correlation coefficient can be shown for the relationship between the response variable and peak_position. This shows that peak_position might have a moderate ability to predict to the weeks_on_chart since there have been

consistently higher correlation coefficients relative to the other independent variables. Additionally, in examining this column, many of the correlation coefficients shown are significantly different from 0 as can be seen from their respective p-values. This differs from the initial model where there were more coefficients that did not show significant correlation between the response variable and the independent variables in the model. Many of these independent variables that showed low correlation coefficients were removed through the iterations we went through to improve our model.

Next, we examined the rest of the correlation matrix to identify coefficients that signify high correlation between certain independent variables. The relationship between [D_p, D_h] remained the highest with a value of -0.555. The relationship between [D_s, D_ne] increased with a correlation coefficient above |0.5| and the correlation coefficient for [D_s, D_c] was shown to be slightly higher in the final model than the initial model with a value above |0.4|. In context, these correlations make sense since country music is expected to be more popular in the south and the audience for pop and hip hop/rap can be expected to be similar. These three relationships were the only ones to show a correlation coefficient above |0.4| and additionally, none of the relationships seem to exhibit severe multicollinearity similarly to the initial model (no coefficients above |0.9|).

As stated in the beginning of this analysis, 50 observations were reserved to be used as test data for the analysis of each residual. As explained above, since the last iteration developed was the "best" model for the data set it was created with, we choose to use that model in order to test and compare its prediction accuracy.

Using the final model, we predicted the value of two observations outside of the data used to create the model but still ensuring to avoid extrapolation:

$\hat{y}^*_{157} = 2.70 - 0.27(3) + 0.76(0) - 0.59(1) -0.10(0) + 0(0) + 0.28(1) + 0.35(0) + 0.95(0) + 0.95(0) + 0.01(-5) = 1.25$

$e^{\hat{y}^*157} = 3.49$

$y_{157} = 24$

$y_{157} - \hat{y}_{157} = 24 - 3.49 = 20.51$

$\hat{y}^*_{172} = 2.70 - 0.27(4) + 0.76(0) - 0.59(0) -0.10(1) + 0(0) + 0.28(0) + 0.35() + 0.95(1) + 0.95(0) + 0.01(-1) = 2.46$

$e^{\hat{y}^*172} = 11.70$

$y_{172} = 52$

$y_{172} - \hat{y}_{172} = 52 - 11.70 = 40.3$

For comparison, we also predicted the value of the same two observations with the initial model:

$\hat{y}_{157}$ = 1.552e01 - 2.14e-01(3) - 9.205e-06(124056) - 4.361e00(0) - 1.007e-01(1) - 3.198e-01(1) + 3.196e00(1) + 1.057e01(0) + 1.250e-02(1) + 9.885e00(0) + 5.851e00(0) -2.532e-01(1) -1.653e00(0) +2.969e00(0) + 4.217e00(0) +3.029e01(0) -8.994e04(119.889) = -61700.015

$y_{157} = 24$

$y_{157} - \hat{y}_{157} = 24 - (-61700.015) = 61724.01$

$\hat{y}_{172}$ = 1.552e01 - 2.14e-01(4) - 9.205e-06(204347) - 4.361e00(0) - 1.007e-01(1) - 3.198e-01(4) + 3.196e00(1) + 1.057e01(0) + 1.250e-02(0) + 9.885e00(1) + 5.851e00(0) -2.532e-01(0) -1.653e00(1) +2.969e00(0) + 4.217e00(0) +3.029e01(0) -8.994e04(124.949) = -66012.07

$y_{172} = 52$

$y_{172} - \hat{y}_{172} = 52 - (-66012.07) = 66064.07$

In context, the final model estimated the number of weeks that the songs Gucci Gang and Believer spent on the Billboard Hot 100 chart to be 3.49 weeks and 11.70 weeks respectively (compared to actual values of 24 and 52). The initial model produced much more inaccurate estimates with values of -61700.015 and -66012.07. This shows that the residual error decreased drastically between the initial and final iterations.

Despite the improvement made through multiple iterations of regression analysis, the final model is still very insufficient in context. The group could have explored other variables that could have made the model better such as the number of streams associated with the song, how many Grammy awards the artist has won, etc. However, it is important to note that the success of a Billboard 100 song is very difficult to predict. It is reasonable to conclude that the length of time that a song spends on the chart is not feasible through the use of simple quantitative variables. Qualitative research would be very important to take into account for an analysis of this scope within the future.

For example, the song called Old Town Road has reached two weeks of Billboard Hot 100 Chart success and it was unexpected by most fans of music within the Hip Hop and Country industries. This song has been a cross-genre success and it was very unpredicted. This is a recent example that shows how the

success of a song is not easy to predict due to unexplainable phenomena. This point led the group to recommend the execution of qualitative surveys presented to fans that exist with each of the industries that have produced chart success to better understand how success occurs on these charts. We believe that surveys will be a better indicator of public interest behind songs.

6. Appendix

Spreadsheets

Final data:
https://docs.google.com/spreadsheets/d/1GQ2vPzBWbJzp0mPJyZYToZmSBJQKtCvTmPaujKqlAHw/edit?usp=sharing

Data with influential points and unusual observations removed:
https://docs.google.com/spreadsheets/d/1GQ2vPzBWbJzp0mPJyZYToZmSBJQKtCvTmPaujKqlAHw/edit?usp=sharing

Binary Variable Coding for Genre and Artist Origination

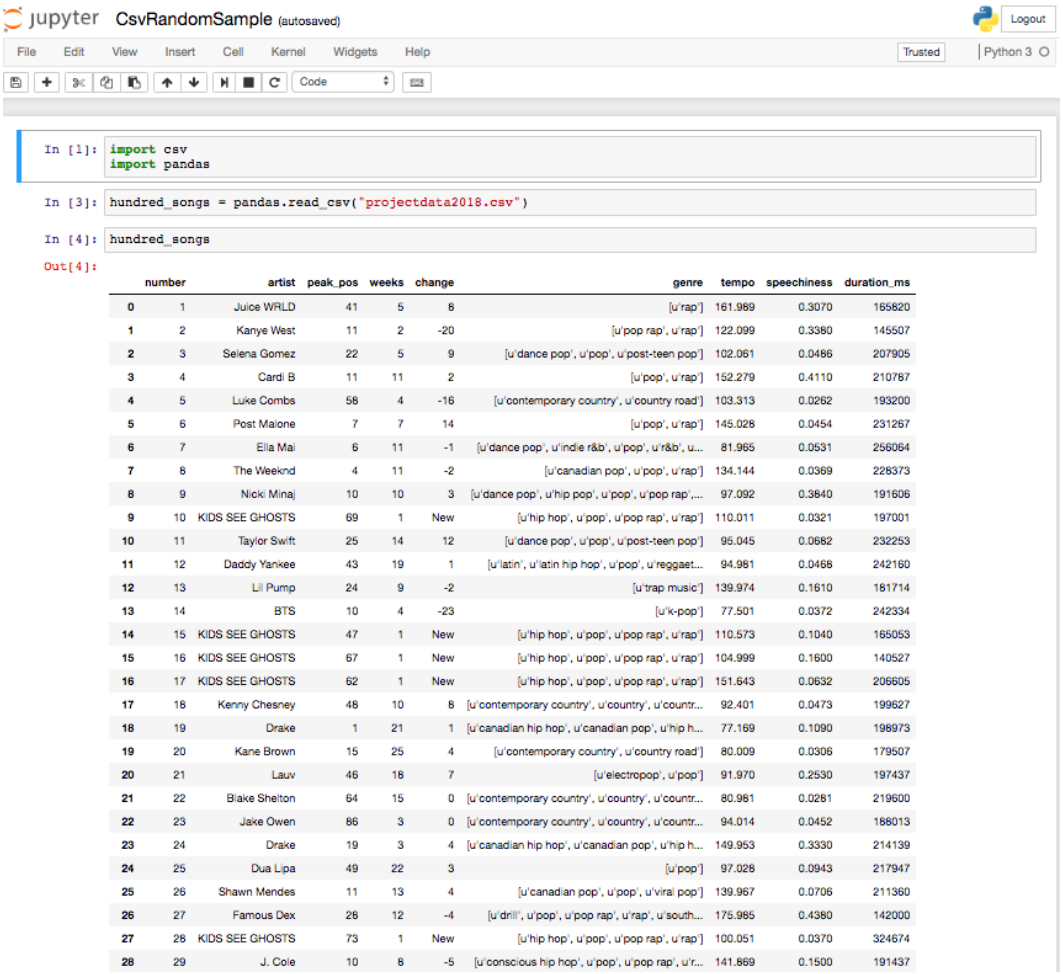
Genre
Base level: R&B

D_c	1 if genre is country 0 otherwise
D_h	1 if genre is country 0 otherwise
D_m	1 if genre is country 0 otherwise
D_p	1 if genre is country 0 otherwise

Artist Organization
Base level: International

D_s	1 if state is in the South 0 otherwise
D_ne	1 if state is in the Northeast 0 otherwise
D_mw	1 if state is in the Midwest 0 otherwise
D_w	1 if state is in the West 0 otherwise
D_pa	1 if state is in the Pacific 0 otherwise

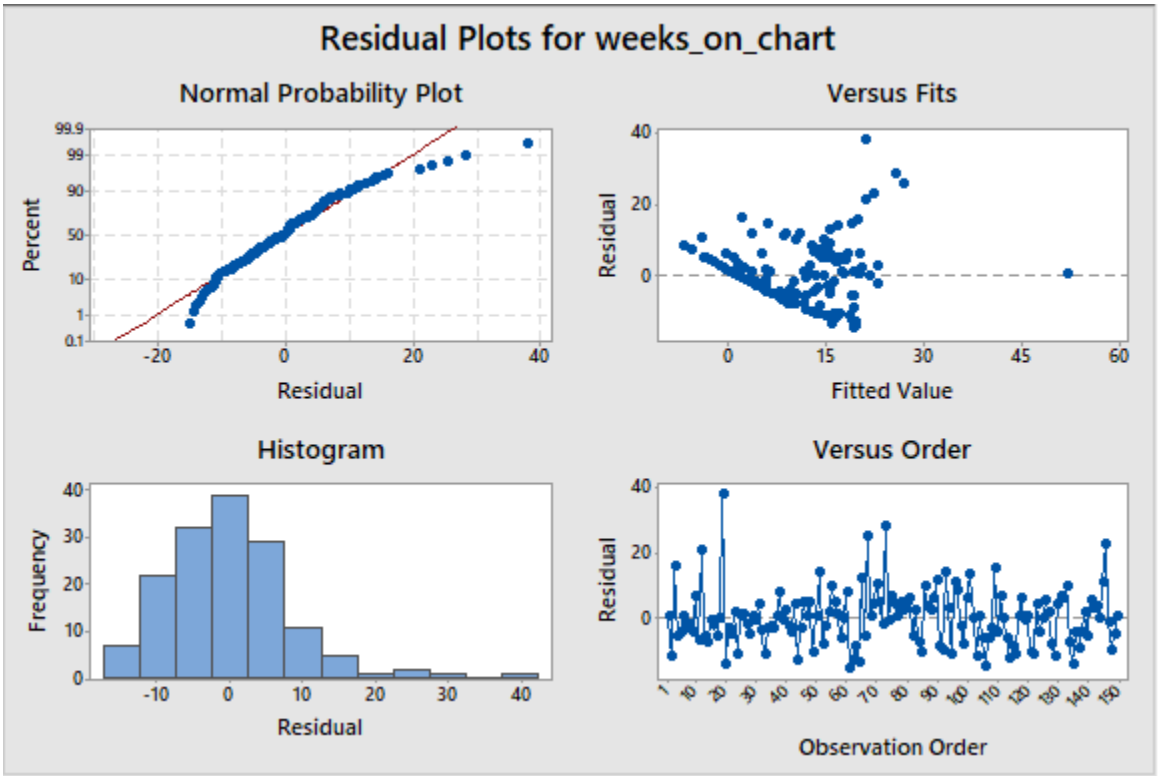
Python Code to randomize 150 observations chosen for model



	number	artist	peak_pos	weeks	change	genre	tempo	speechiness	duration_ms
0	1	Juice WRLD	41	5	8	[u'rap']	161.989	0.3070	165820
1	2	Kanye West	11	2	-20	[u'pop rap', u'rap']	122.099	0.3380	145507
2	3	Selena Gomez	22	5	9	[u'dance pop', u'pop', u'post-teen pop']	102.061	0.0486	207905
3	4	Cardi B	11	11	2	[u'pop', u'rap']	152.279	0.4110	210787
4	5	Luke Combs	58	4	-16	[u'contemporary country', u'country road']	103.313	0.0262	193200
5	6	Post Malone	7	7	14	[u'pop', u'rap']	145.028	0.0454	231267
6	7	Ella Mai	6	11	-1	[u'dance pop', u'indie r&b', u'pop', u'r&b', u...	81.965	0.0531	256064
7	8	The Weeknd	4	11	-2	[u'canadian pop', u'pop', u'rap']	134.144	0.0369	228373
8	9	Nicki Minaj	10	10	3	[u'dance pop', u'hip pop', u'pop', u'pop rap',...	97.092	0.3840	191606
9	10	KIDS SEE GHOSTS	69	1	New	[u'hip hop', u'pop', u'pop rap', u'rap']	110.011	0.0321	197001
10	11	Taylor Swift	25	14	12	[u'dance pop', u'pop', u'post-teen pop']	95.045	0.0682	232253
11	12	Daddy Yankee	43	19	1	[u'latin', u'latin hip hop', u'pop', u'reggaet...	94.981	0.0468	242160
12	13	Lil Pump	24	9	-2	[u'trap music']	139.974	0.1610	181714
13	14	BTS	10	4	-23	[u'k-pop']	77.501	0.0372	242334
14	15	KIDS SEE GHOSTS	47	1	New	[u'hip hop', u'pop', u'pop rap', u'rap']	110.573	0.1040	165053
15	16	KIDS SEE GHOSTS	67	1	New	[u'hip hop', u'pop', u'pop rap', u'rap']	104.999	0.1600	140527
16	17	KIDS SEE GHOSTS	62	1	New	[u'hip hop', u'pop', u'pop rap', u'rap']	151.643	0.0632	206605
17	18	Kenny Chesney	48	10	8	[u'contemporary country', u'country', u'countr...	92.401	0.0473	199627
18	19	Drake	1	21	1	[u'canadian hip hop', u'canadian pop', u'hip h...	77.169	0.1090	198973
19	20	Kane Brown	15	25	4	[u'contemporary country', u'country road']	80.009	0.0306	179507
20	21	Lauv	46	18	7	[u'electropop', u'pop']	91.970	0.2530	197437
21	22	Blake Shelton	64	15	0	[u'contemporary country', u'country', u'countr...	80.981	0.0281	219600
22	23	Jake Owen	86	3	0	[u'contemporary country', u'country', u'countr...	94.014	0.0452	188013
23	24	Drake	19	3	4	[u'canadian hip hop', u'canadian pop', u'hip h...	149.953	0.3330	214139
24	25	Dua Lipa	49	22	3	[u'pop']	97.028	0.0943	217947
25	26	Shawn Mendes	11	13	4	[u'canadian pop', u'pop', u'viral pop']	139.967	0.0706	211360
26	27	Famous Dex	28	12	-4	[u'drill', u'pop', u'pop rap', u'rap', u'south...	175.985	0.4380	142000
27	28	KIDS SEE GHOSTS	73	1	New	[u'hip hop', u'pop', u'pop rap', u'rap']	100.051	0.0370	324674
28	29	J. Cole	10	8	-5	[u'conscious hip hop', u'pop', u'pop rap', u'r...	141.869	0.1500	191437

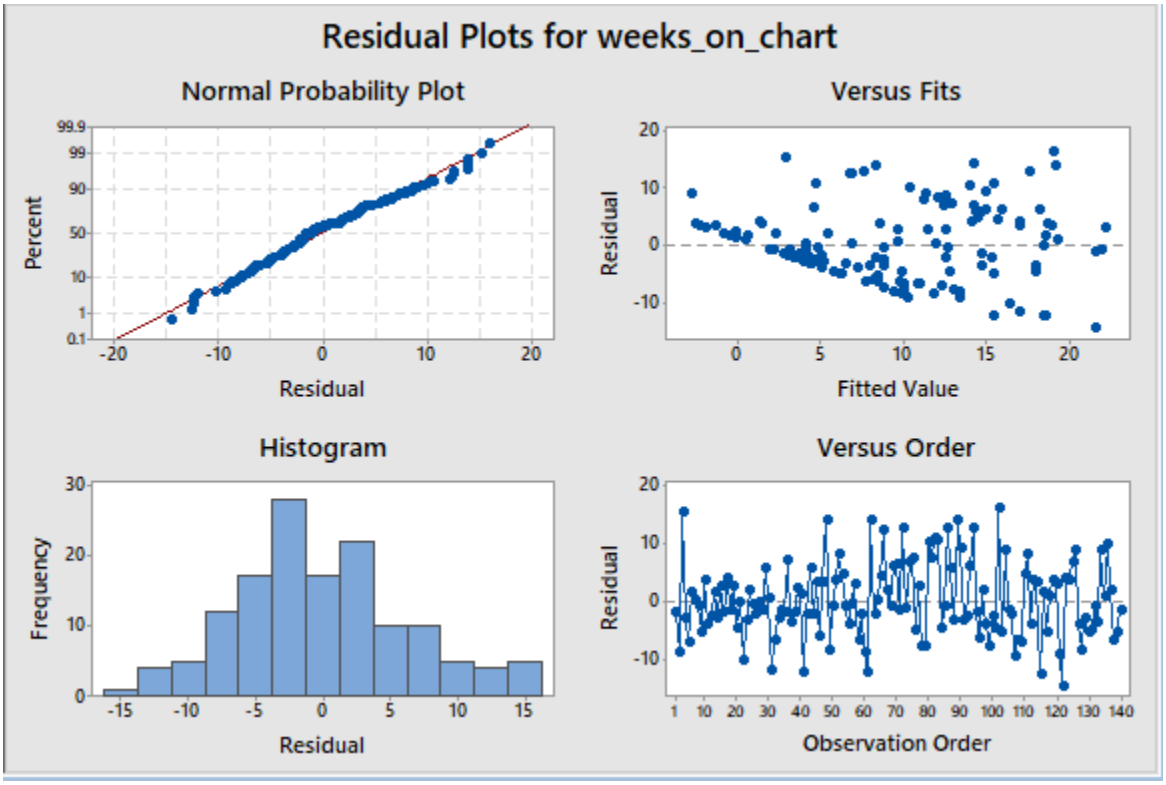
```
In [6]: sample_150 = hundred_songs.sample(n = 150)

In [9]: list(sample_150["number"])

Out[9]: [145,
         82,
         127,
         117,
         1,
         116,
         103,
         148,
         150,
         91,
         136,
         40,
         36,
         48,
         17,
         55,
         169,
         21,
         160,
         124,
         204,

In [12]: a = {}
         for item in list(sample_150["number"]):
             if str(item) in a:
                 a[str(item)]+=1
             else:
                 a[str(item)]= 1
         a

Out[12]: {'1': 1,
          '10': 1,
          '100': 1,
          '101': 1,
          '102': 1,
          '103': 1,
          '104': 1,
          '106': 1,
          '108': 1,
```

Residual Analysis - All variables, first model



Mean	-7.50511E-16
StDev	8.670
N	150
AD	1.210
P-Value	<0.005

Model without unusual observations



Mean	8.945226E-16
StDev	6.390
N	140
AD	0.756
P-Value	0.048

Transformed ln(y)

Regression Analysis: ln(y) versus peak_position, change, ... D_mw, D_w

The following terms cannot be estimated and were removed:
D_p

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	9	115.926	12.8807	18.06	0.000
peak_position	1	73.298	73.2975	102.77	0.000
change	1	6.988	6.9877	9.80	0.002
D_c	1	7.668	7.6681	10.75	0.001
D_h	1	7.641	7.6413	10.71	0.001
D_m	1	0.086	0.0857	0.12	0.729
D_s	1	0.945	0.9454	1.33	0.252
D_ne	1	1.377	1.3772	1.93	0.167
D_mw	1	5.283	5.2827	7.41	0.007
D_w	1	7.546	7.5460	10.58	0.001
Error	130	92.716	0.7132		
Lack-of-Fit	126	92.441	0.7337	10.66	0.016
Pure Error	4	0.275	0.0688		
Total	139	208.642			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.844512	55.56%	52.49%	49.00%

Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	2.745	0.235	(2.281, 3.209)	11.70	0.000	
peak_position	-0.02774	0.00274	(-0.03316, -0.02233)	-10.14	0.000	1.25
change	0.01632	0.00521	(0.00600, 0.02663)	3.13	0.002	1.11
D_c	0.763	0.233	(0.303, 1.224)	3.28	0.001	1.75
D_h	-0.599	0.183	(-0.961, -0.237)	-3.27	0.001	1.55
D_m	-0.104	0.299	(-0.695, 0.488)	-0.35	0.729	1.27
D_s	0.284	0.247	(-0.204, 0.772)	1.15	0.252	2.97
D_ne	0.355	0.255	(-0.150, 0.860)	1.39	0.167	2.16
D_mw	0.950	0.349	(0.260, 1.641)	2.72	0.007	1.44
D_w	0.955	0.294	(0.374, 1.536)	3.25	0.001	1.90

Regression Equation

ln(y) = 2.745 - 0.02774 peak_position + 0.01632 change + 0.763 D_c - 0.599 D_h - 0.104 D_m + 0.284 D_s + 0.355 D_ne + 0.950 D_mw + 0.955 D_w

Fits and Diagnostics for Unusual Observations

Obs	ln(y)	Fit	SE Fit	95% CI	Resid	Std Resid	Del Resid	HI	Cook's D
3	2.890	0.607	0.141	(0.328, 0.886)	2.283	2.74	2.81	0.0278614	0.02
83	2.708	0.789	0.201	(0.392, 1.185)	1.919	2.34	2.38	0.0563854	0.03
86	2.996	1.309	0.237	(0.839, 1.779)	1.687	2.08	2.11	0.0790742	0.04
89	3.091	1.434	0.226	(0.987, 1.882)	1.657	2.04	2.06	0.0717982	0.03
126	1.792	-0.230	0.171	(-0.569, 0.108)	2.022	2.44	2.49	0.0409724	0.03

Obs	DFITS	
3	0.476458	R
83	0.582093	R
86	0.617935	R
89	0.573369	R
126	0.515411	R

R Large residual

Durbin-Watson Statistic

Durbin-Watson Statistic = 2.05077

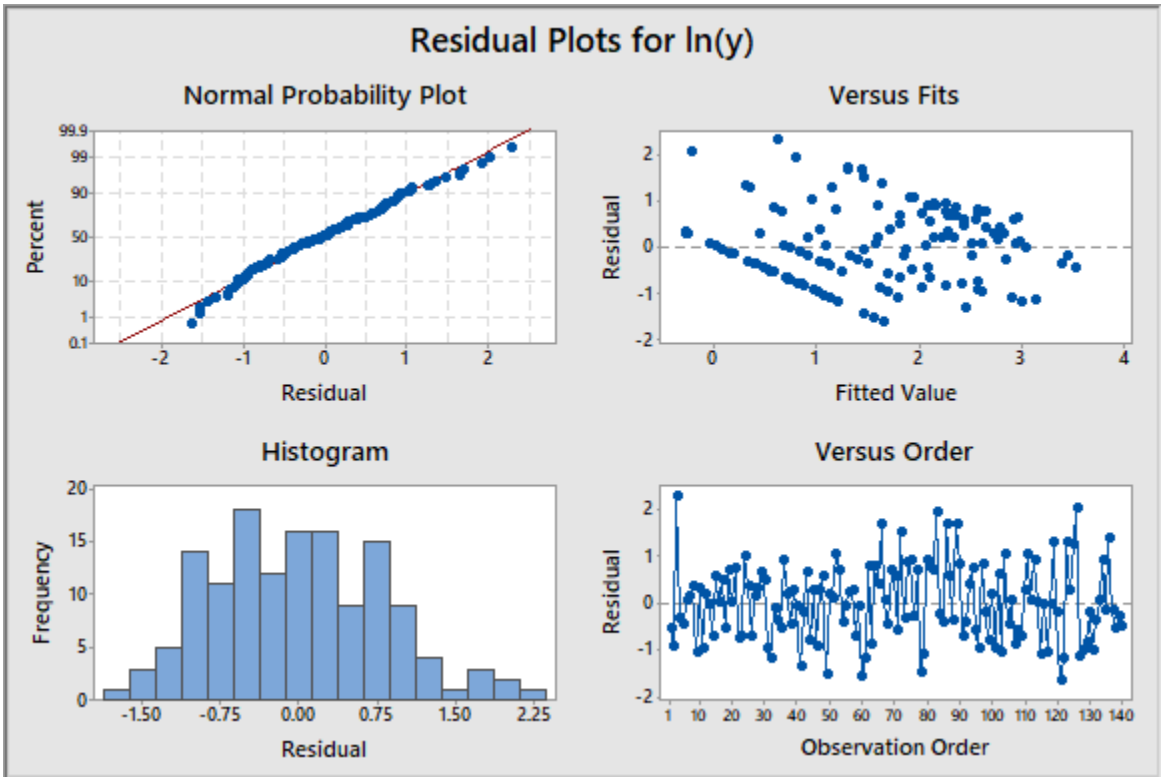
Correlation: ln(y), D_c, D_h, D_m, D_p, D_s, D_ne, D_mw, ... ion, change

Correlations

	ln(y)	D_c	D_h	D_m	D_p
D_c	0.222				
	0.008				
D_h	-0.350	-0.399			
	0.000	0.000			
D_m	-0.021	-0.149	-0.228		
	0.807	0.078	0.007		
D_p	0.181	-0.363	-0.555	-0.208	
	0.033	0.000	0.000	0.014	
D_s	-0.176	0.408	0.100	-0.165	-0.359
	0.037	0.000	0.239	0.051	0.000
D_ne	0.047	-0.267	-0.120	-0.023	0.366
	0.582	0.001	0.156	0.786	0.000
D_mw	0.110	-0.062	0.096	0.032	-0.063
	0.194	0.466	0.261	0.710	0.460
D_w	0.133	-0.091	-0.080	0.284	-0.002
	0.118	0.285	0.349	0.001	0.982
peak_position	-0.539	0.191	-0.021	0.141	-0.222
	0.000	0.024	0.801	0.095	0.008
change	0.091	0.086	-0.127	0.108	-0.005
	0.285	0.312	0.135	0.204	0.954

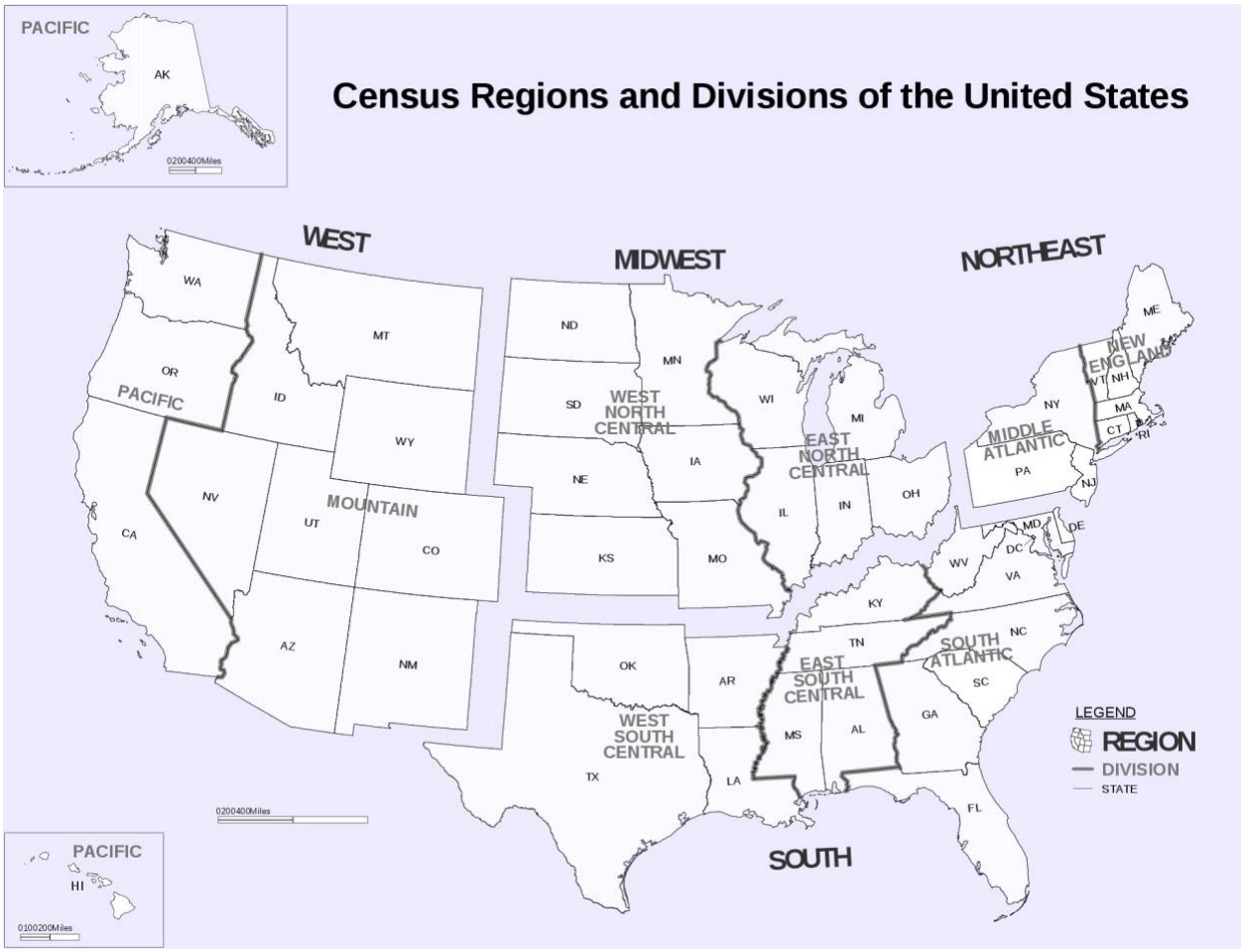
	D_s	D_ne	D_mw	D_w	peak_position
D_ne	-0.486				
	0.000				
D_mw	-0.244	-0.137			
	0.004	0.107			
D_w	-0.358	-0.201	-0.101		
	0.000	0.017	0.237		
peak_position	0.273	-0.186	-0.039	0.084	
	0.001	0.028	0.644	0.323	
change	0.018	-0.115	0.018	0.114	0.261
	0.833	0.174	0.833	0.181	0.002

Transformed model



Mean	1.256931E-15
StDev	0.8167
N	140
AD	0.419
P-Value	0.322

Map used to re-code artist origination variable from 50 US states to regions within the US

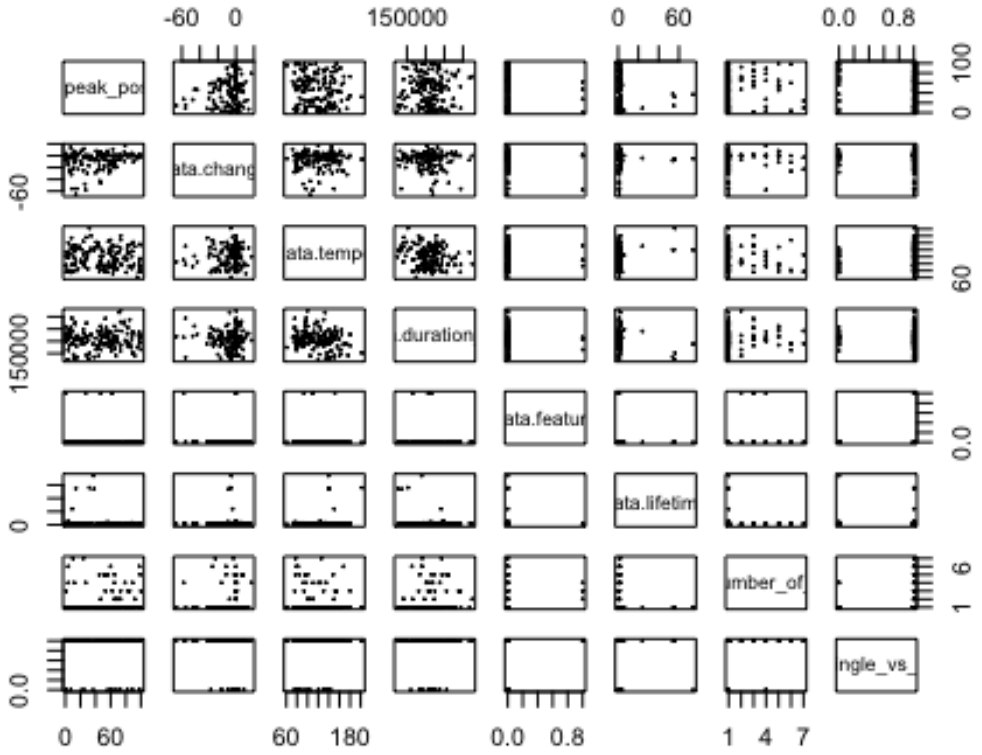


R code consisting of graphs, charts and code chunks

The following code chunk attaches the data set and makes a matrix of independent variables to show their correlation.

```
library(readxl)
data <- read_excel("FinalProjectData.xlsx")
quantvars <-read_excel("quantvars.xlsx")
testingdata <-read_excel("rdata.xlsx")
attach(data)
attach(quantvars)

## The following objects are masked from data:
##
##     artist_origination, Genre, number

attach(testingdata)
```

```
## The following objects are masked from quantvars:
##
##     D_c, D_h, D_m, D_mw, D_ne, D_p, D_pa, D_s, D_w, number

## The following objects are masked from data:
##
##     artist, change, duration_ms, feature, lifetime, number,
##     number_of_artists, peak_position, single_vs_album, song_name,
##     tempo, weeks_on_chart

quantdata <- data.frame(data$peak_position,data$change,data$tempo,data$duration_ms,data$feature,dat
a$lifetime,data$number_of_artists,data$single_vs_album) #gets quant variables
pairs(quantdata,pch=21,cex=0.1) #makes matrix plot
```

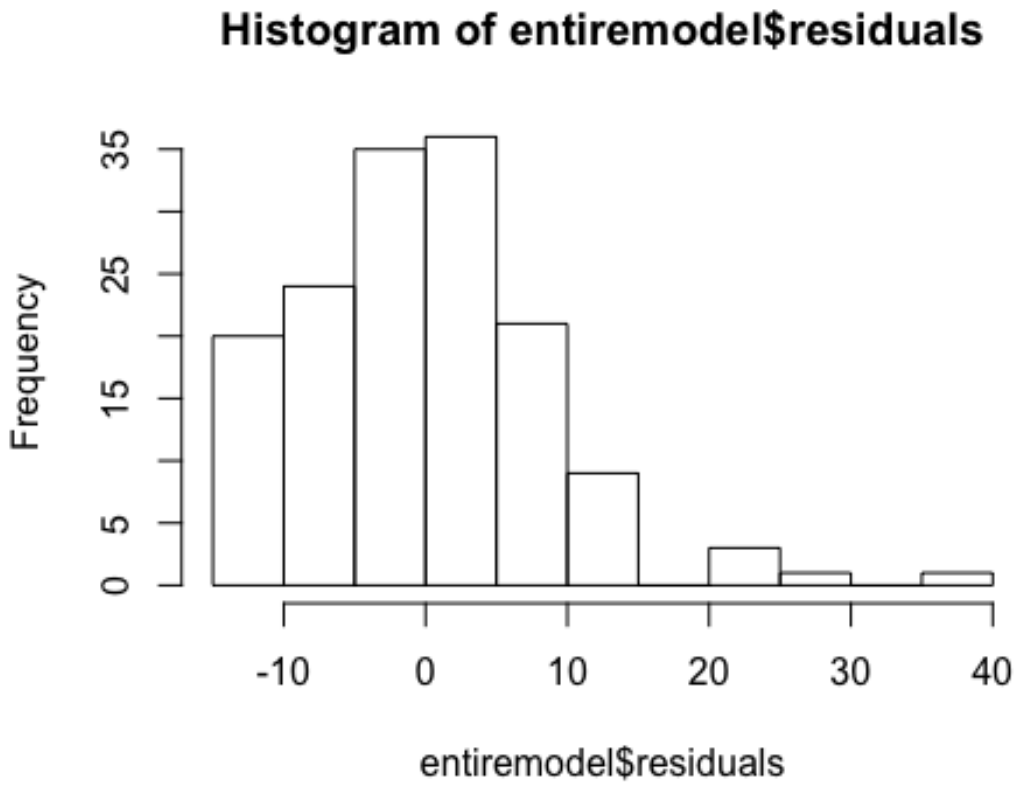


The following code chunk creates vectors of the variables and then runs a MLR model of all the possible independent variables that could predict the number of weeks a song remains on the chart. It also analyses the assumptions of the residuals of the model created (histogram, QQplot and scatterplot).

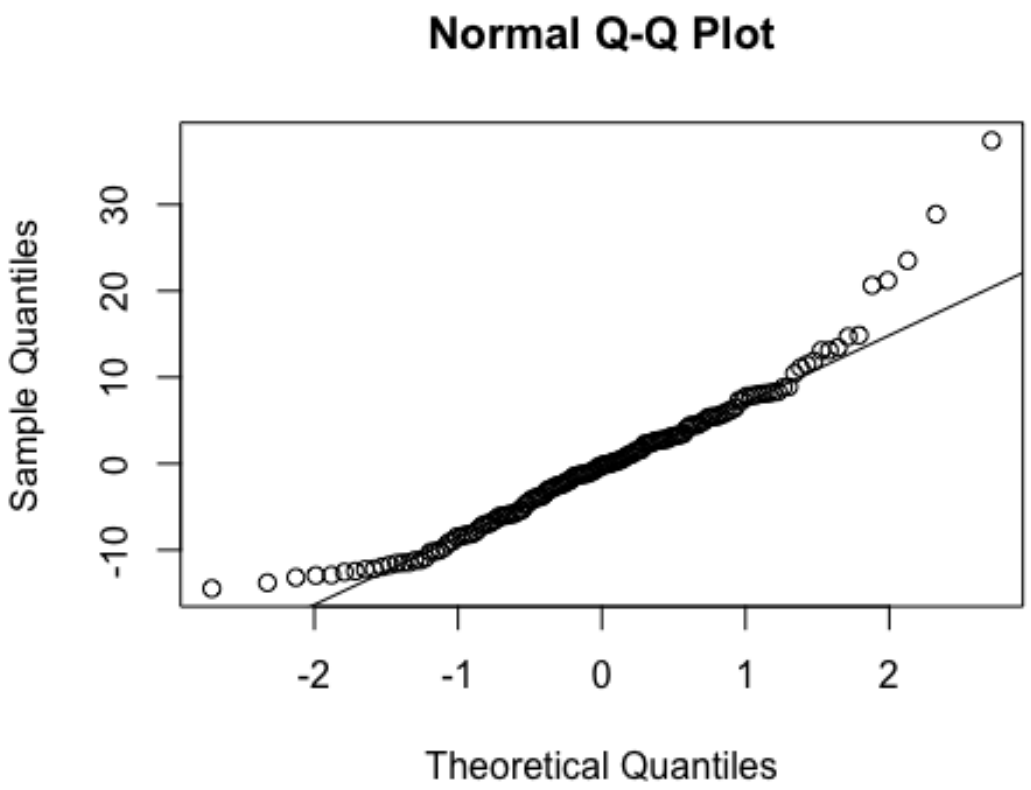
```
y=data$weeks_on_chart
position = data$peak_position
change = data$change
tempo = data$tempo
duration = data$duration_ms
feature = data$feature
```

```
lifetime = data$lifetime
numartists = data$number_of_artists
singleoralbum = data$single_vs_album
#dummy variables for genre
country = quantvars$D_c
hiprap = quantvars$D_h
misc = quantvars$D_m
pop = quantvars$D_p
#dummy variables for artist origination
south = quantvars$D_s
northeast = quantvars$D_ne
midwest = quantvars$D_mw
west = quantvars$D_w
pacific = quantvars$D_pa
entiremodel = lm(y~position+duration+feature+lifetime+numartists+singleoralbum+country+hiprap+mis
c+pop+south+northeast+midwest+west+pacific+tempo)
summary(entiremodel)

##
## Call:
## lm(formula = y ~ position + duration + feature + lifetime + numartists +
##     singleoralbum + country + hiprap + misc + pop + south + northeast +
##     midwest + west + pacific + tempo)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.467  -6.002  -0.243   4.517  37.404
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.552e+01  8.685e+00   1.787  0.07618 .
## position      -2.140e-01  2.778e-02  -7.702 2.72e-12 ***
## duration      -9.205e-06  2.193e-05  -0.420  0.67530
## feature       -4.361e+00  4.934e+00  -0.884  0.37835
## lifetime      -1.007e-01  9.341e-02  -1.078  0.28300
## numartists    -3.198e-01  6.649e-01  -0.481  0.63133
## singleoralbum  3.196e+00  2.534e+00   1.261  0.20943
## country        1.057e+01  5.233e+00   2.020  0.04544 *
## hiprap         1.250e-02  4.989e+00   0.003  0.99801
## misc           9.885e+00  5.828e+00   1.696  0.09217 .
## pop            5.851e+00  4.991e+00   1.172  0.24315
## south         -2.532e-01  2.588e+00  -0.098  0.92222
## northeast     -1.653e+00  2.658e+00  -0.622  0.53503
## midwest        2.969e+00  3.658e+00   0.812  0.41852
## west           4.217e+00  3.138e+00   1.344  0.18139
## pacific        3.029e+01  9.339e+00   3.243  0.00149 **
## tempo         -8.994e-04  2.948e-02  -0.031  0.97571
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.036 on 133 degrees of freedom
```

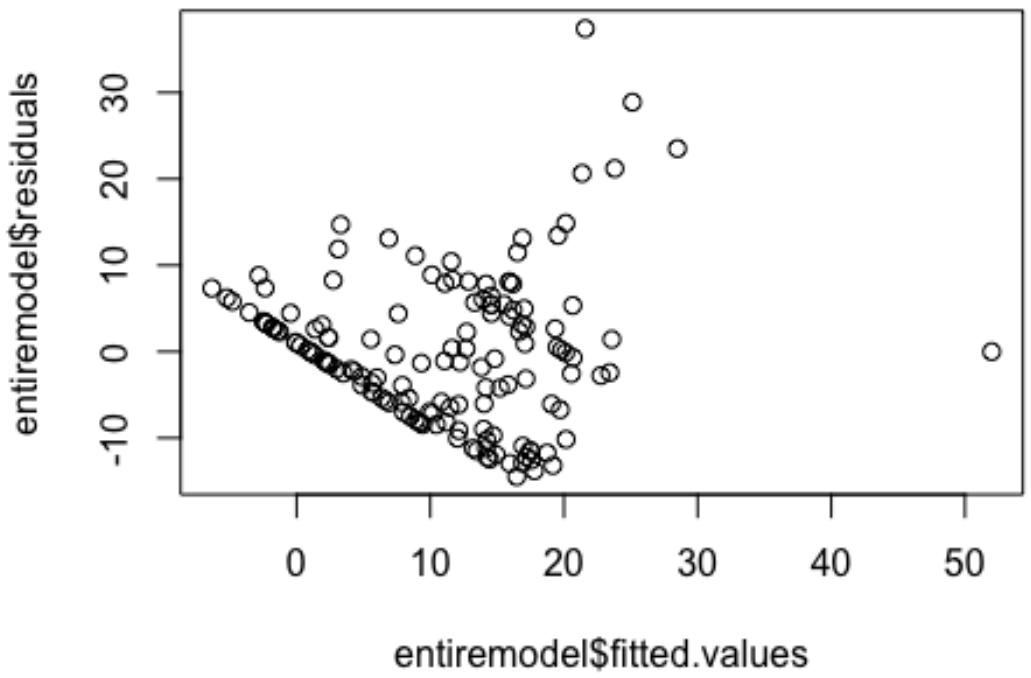
```
## Multiple R-squared:  0.4893, Adjusted R-squared:  0.4278
## F-statistic: 7.963 on 16 and 133 DF,  p-value: 5.059e-13

#Assumption 1: E[residuals] = 0
hist(entiremodel$residuals) #histogram
```



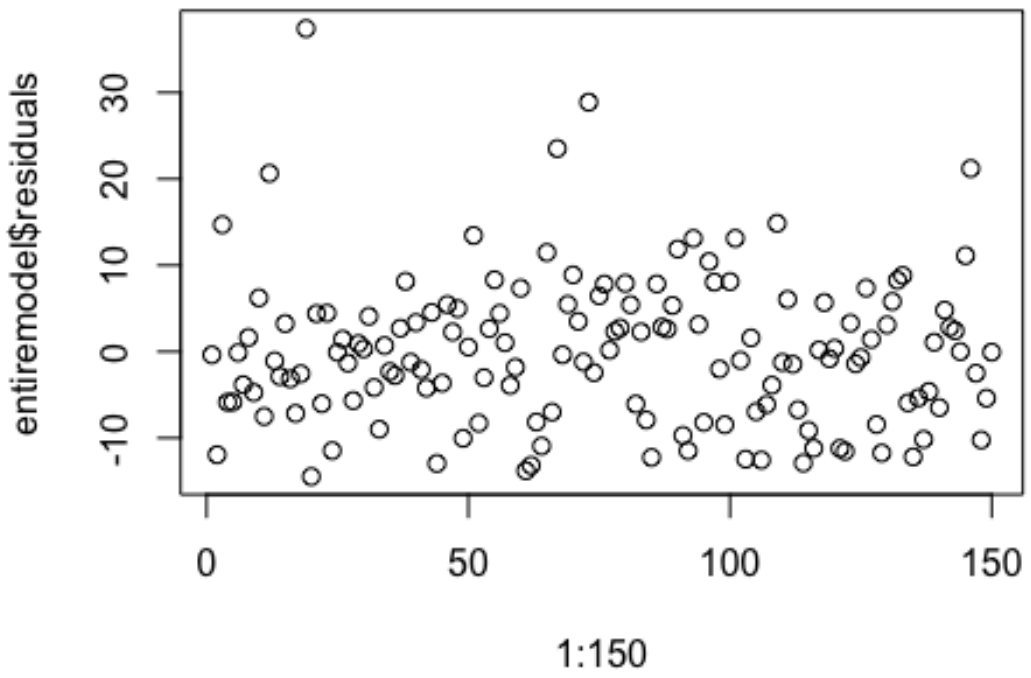

```
mean(entiremodel$residuals) #mean of residuals

## [1] -1.987761e-16

qqnorm(entiremodel$residuals)
qqline(entiremodel$residuals)
```




```
#Assumption 2: Normality
shapiro.test(entiremodel$residuals)

##
##  Shapiro-Wilk normality test
##
## data:  entiremodel$residuals
## W = 0.94548, p-value = 1.405e-05

#Assumption 3: Identical (constant) variance
plot(entiremodel$fitted.values,entiremodel$residuals)
```




```
plot(1:150,entiremodel$residuals)
```



This code chunk is an ANOVA table for the model with all possible variables in it. It also has different screening methods: stepwise regression, backwards elimination and best subsets.

```
anova(entiremodel)

## Analysis of Variance Table
##
## Response: y
##               Df  Sum Sq Mean Sq F value    Pr(>F)
## position       1  6000.7  6000.7 73.4974  2.25e-14 ***
## duration       1     3.3     3.3  0.0407 0.8404084
## feature        1   406.5   406.5  4.9785 0.0273393 *
## lifetime       1     0.2     0.2  0.0021 0.9637424
## numartists     1   125.6   125.6  1.5389 0.2169721
## singleoralbum  1   172.7   172.7  2.1153 0.1481864
## country        1  1228.8  1228.8 15.0504 0.0001639 ***
## hiprap         1   853.3   853.3 10.4513 0.0015456 **
## misc           1   201.8   201.8  2.4717 0.1182900
## pop            1    84.0    84.0  1.0291 0.3122177
## south          1    64.1    64.1  0.7850 0.3772060
## northeast      1   310.5   310.5  3.8026 0.0532764 .
## midwest        1     1.8     1.8  0.0219 0.8825617
## west           1    87.3    87.3  1.0695 0.3029428
## pacific        1   861.8   861.8 10.5561 0.0014665 **
```

```
## tempo          1     0.1     0.1  0.0009 0.9757060
## Residuals    133 10858.7    81.6

## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

#Apply the stepwise regression and obtain the solution.
library(MASS)
step<-stepAIC(entiremodel, direction="both")

## Start:  AIC=676.31
## y ~ position + duration + feature + lifetime + numartists + singleoralbum +
##     country + hiprap + misc + pop + south + northeast + midwest +
##     west + pacific + tempo
##
##                Df Sum of Sq    RSS    AIC
## - hiprap        1       0.0 10859 674.31
## - tempo         1       0.1 10859 674.31
## - south         1       0.8 10860 674.32
## - duration      1      14.4 10873 674.51
## - numartists    1      18.9 10878 674.57
## - northeast     1      31.6 10890 674.75
## - midwest       1      53.8 10912 675.05
## - feature       1      63.8 10922 675.19
## - lifetime      1      94.9 10954 675.62
## - pop           1     112.2 10971 675.86
## - singleoralbum 1     129.9 10989 676.10
## <none>                      10859 676.31
## - west          1     147.4 11006 676.34
## - misc          1     234.9 11094 677.52
## - country       1     333.0 11192 678.84
## - pacific       1     858.9 11718 685.73
## - position      1    4843.8 15702 729.64
##
## Step:  AIC=674.31
## y ~ position + duration + feature + lifetime + numartists + singleoralbum +
##     country + misc + pop + south + northeast + midwest + west +
##     pacific + tempo
##
##                Df Sum of Sq    RSS    AIC
## - tempo         1       0.1 10859 672.31
## - south         1       0.8 10860 672.32
## - duration      1      14.6 10873 672.52
## - numartists    1      18.9 10878 672.57
## - northeast     1      31.6 10890 672.75
## - midwest       1      54.8 10914 673.07
## - feature       1      64.0 10923 673.20
## - lifetime      1      95.0 10954 673.62
## - singleoralbum 1     131.5 10990 674.12
## <none>                      10859 674.31
## - west          1     148.5 11007 674.35
```

```
## + hiprap         1       0.0 10859 676.31
## - pop            1     629.7 11488 680.77
## - misc           1     630.8 11490 680.78
## - pacific        1     858.9 11718 683.73
## - country        1    1770.7 12629 694.97
## - position       1    4902.7 15762 728.20
##
## Step:  AIC=672.31
## y ~ position + duration + feature + lifetime + numartists + singleoralbum +
##     country + misc + pop + south + northeast + midwest + west +
##     pacific
##
##                Df Sum of Sq    RSS    AIC
## - south         1       0.9 10860 670.33
## - duration      1      14.6 10873 670.52
## - numartists    1      18.9 10878 670.58
## - northeast     1      33.9 10893 670.78
## - midwest       1      56.5 10915 671.09
## - feature       1      64.0 10923 671.20
## - lifetime      1     101.5 10960 671.71
## - singleoralbum 1     132.8 10992 672.14
## <none>                      10859 672.31
## - west          1     148.9 11008 672.36
## + tempo         1       0.1 10859 674.31
## + hiprap        1       0.0 10859 674.31
## - pop           1     634.8 11494 678.84
## - misc          1     663.3 11522 679.21
## - pacific       1     861.8 11721 681.77
## - country       1    1929.1 12788 694.84
## - position      1    4929.3 15788 726.46
##
## Step:  AIC=670.33
## y ~ position + duration + feature + lifetime + numartists + singleoralbum +
##     country + misc + pop + northeast + midwest + west + pacific
##
##                Df Sum of Sq    RSS    AIC
## - duration      1      14.0 10874 668.52
## - numartists    1      20.0 10880 668.60
## - northeast     1      41.1 10901 668.89
## - feature       1      63.4 10923 669.20
## - midwest       1      92.3 10952 669.60
## - lifetime      1     106.9 10967 669.80
## - singleoralbum 1     132.6 10992 670.15
## <none>                      10860 670.33
## - west          1     247.2 11107 671.70
## + south         1       0.9 10859 672.31
## + tempo         1       0.2 10860 672.32
## + hiprap        1       0.0 10860 672.33
## - pop           1     664.0 11524 677.23
## - misc          1     679.7 11540 677.43
## - pacific       1     880.4 11740 680.02
```

```
## - country        1    1963.3 12823 693.25
## - position       1    5531.6 16391 730.08
##
## Step:  AIC=668.52
## y ~ position + feature + lifetime + numartists + singleoralbum +
##     country + misc + pop + northeast + midwest + west + pacific
##
##                Df Sum of Sq   RSS    AIC
## - numartists    1      21.0 10895 666.81
## - northeast     1      46.6 10920 667.16
## - feature       1      63.5 10937 667.39
## - lifetime      1      95.4 10969 667.83
## - midwest       1      96.5 10970 667.84
## - singleoralbum 1     143.5 11017 668.49
## <none>                      10874 668.52
## - west          1     245.4 11119 669.87
## + duration      1      14.0 10860 670.33
## + south         1       0.2 10873 670.52
## + hiprap        1       0.2 10874 670.52
## + tempo         1       0.0 10874 670.52
## - pop           1     659.6 11533 675.35
## - misc          1     669.3 11543 675.48
## - pacific       1     885.9 11760 678.27
## - country       1    1949.9 12824 691.26
## - position      1    5563.6 16437 728.50
##
## Step:  AIC=666.81
## y ~ position + feature + lifetime + singleoralbum + country +
##     misc + pop + northeast + midwest + west + pacific
##
##                Df Sum of Sq   RSS    AIC
## - northeast     1      35.3 10930 665.30
## - lifetime      1      79.7 10974 665.90
## - feature       1      90.6 10985 666.05
## - midwest       1      96.2 10991 666.13
## - singleoralbum 1     134.9 11030 666.66
## <none>                      10895 666.81
## - west          1     224.9 11120 667.88
## + numartists    1      21.0 10874 668.52
## + duration      1      15.0 10880 668.60
## + south         1       0.9 10894 668.80
## + hiprap        1       0.1 10895 668.81
## + tempo         1       0.0 10895 668.81
## - pop           1     644.2 11539 673.43
## - misc          1     672.8 11568 673.80
## - pacific       1     906.2 11801 676.79
## - country       1    1952.4 12847 689.54
## - position      1    5646.4 16541 727.45
##
## Step:  AIC=665.3
## y ~ position + feature + lifetime + singleoralbum + country +
```

```
##     misc + pop + midwest + west + pacific
##
##                Df Sum of Sq   RSS    AIC
## - lifetime      1      71.8 11002 664.28
## - feature       1      91.6 11022 664.55
## - midwest       1     118.9 11049 664.92
## - singleoralbum 1     126.7 11057 665.02
## <none>                      10930 665.30
## + northeast     1      35.3 10895 666.81
## + duration      1      19.7 10910 667.03
## + numartists    1       9.8 10920 667.16
## + south         1       7.7 10922 667.19
## - west          1     291.8 11222 667.25
## + hiprap        1       0.6 10930 667.29
## + tempo         1       0.5 10930 667.29
## - pop           1     609.2 11539 671.43
## - misc          1     648.6 11579 671.94
## - pacific       1     955.8 11886 675.87
## - country       1    2107.5 13038 689.74
## - position      1    5618.5 16549 725.51
##
## Step:  AIC=664.28
## y ~ position + feature + singleoralbum + country + misc + pop +
##     midwest + west + pacific
##
##                Df Sum of Sq   RSS    AIC
## - feature       1      91.3 11093 663.52
## - midwest       1     100.7 11102 663.64
## - singleoralbum 1     145.7 11148 664.25
## <none>                      11002 664.28
## + lifetime      1      71.8 10930 665.30
## + northeast     1      27.3 10974 665.90
## + tempo         1       9.0 10993 666.15
## + duration      1       5.4 10996 666.20
## + south         1       1.5 11000 666.26
## + numartists    1       1.5 11000 666.26
## + hiprap        1       0.4 11001 666.27
## - west          1     313.8 11316 666.50
## - pop           1     576.3 11578 669.94
## - misc          1     580.1 11582 669.98
## - pacific       1     964.1 11966 674.88
## - country       1    2065.8 13068 688.09
## - position      1    5555.7 16558 723.59
##
## Step:  AIC=663.52
## y ~ position + singleoralbum + country + misc + pop + midwest +
##     west + pacific
##
##                Df Sum of Sq   RSS    AIC
## - midwest       1     117.4 11210 663.10
## - singleoralbum 1     131.8 11225 663.29
```

```
## <none>                     11093 663.52
## + feature       1      91.3 11002 664.28
## + lifetime      1      71.5 11022 664.55
## + northeast     1      28.2 11065 665.13
## + numartists    1      12.7 11080 665.34
## + duration      1       5.9 11087 665.44
## + tempo         1       5.6 11088 665.44
## + south         1       2.1 11091 665.49
## + hiprap        1       0.0 11093 665.52
## - west          1     331.7 11425 665.94
## - misc          1     629.7 11723 669.80
## - pop           1     705.1 11798 670.76
## - pacific       1     970.4 12064 674.10
## - country       1    2272.7 13366 689.47
## - position      1    5484.0 16577 721.77
##
## Step:  AIC=663.1
## y ~ position + singleoralbum + country + misc + pop + west +
##     pacific
##
##                Df Sum of Sq    RSS    AIC
## - singleoralbum  1     120.0 11330 662.69
## <none>                      11210 663.10
## + midwest       1     117.4 11093 663.52
## + feature       1     108.0 11102 663.64
## + lifetime      1      52.0 11158 664.40
## + northeast     1      50.2 11160 664.42
## + numartists    1      13.2 11197 664.92
## - west          1     290.7 11501 664.94
## + duration      1      11.9 11199 664.94
## + south         1       7.2 11203 665.00
## + hiprap        1       6.8 11204 665.00
## + tempo         1       0.3 11210 665.09
## - misc          1     621.5 11832 669.19
## - pop           1     650.3 11861 669.55
## - pacific       1     962.9 12173 673.46
## - country       1    2177.1 13388 687.72
## - position      1    5421.2 16632 720.26
##
## Step:  AIC=662.69
## y ~ position + country + misc + pop + west + pacific
##
##                Df Sum of Sq    RSS    AIC
## <none>                      11330 662.69
## + singleoralbum  1     120.0 11210 663.10
## + midwest       1     105.5 11225 663.29
## + feature       1      92.6 11238 663.46
## + lifetime      1      67.6 11263 663.79
## + northeast     1      38.3 11292 664.19
## - west          1     281.5 11612 664.37
## + duration      1      17.6 11313 664.46
```

```
## + hiprap         1      12.6 11318 664.53
## + south          1       7.8 11323 664.59
## + numartists     1       6.1 11324 664.61
## + tempo          1       0.4 11330 664.69
## - misc           1     589.9 11920 668.31
## - pop            1     728.6 12059 670.04
## - pacific        1     961.5 12292 672.91
## - country        1    2247.9 13578 687.84
## - position       1    5696.3 17027 721.79

step$anova # display results

## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## y ~ position + duration + feature + lifetime + numartists + singleoralbum +
##     country + hiprap + misc + pop + south + northeast + midwest +
##     west + pacific + tempo
##
## Final Model:
## y ~ position + country + misc + pop + west + pacific
##
##
##              Step Df     Deviance Resid. Df Resid. Dev      AIC
## 1                                       133   10858.73 676.3134
## 2         - hiprap  1 5.121132e-04       134   10858.73 674.3134
## 3          - tempo  1 7.627170e-02       135   10858.81 672.3145
## 4          - south  1 8.998228e-01       136   10859.71 670.3269
## 5       - duration  1 1.396770e+01       137   10873.67 668.5197
## 6     - numartists  1 2.104988e+01       138   10894.72 666.8098
## 7      - northeast  1 3.531772e+01       139   10930.04 665.2953
## 8       - lifetime  1 7.175365e+01       140   11001.80 664.2768
## 9        - feature  1 9.134956e+01       141   11093.15 663.5171
## 10       - midwest  1 1.173522e+02       142   11210.50 663.0956
## 11 - singleoralbum  1 1.200187e+02       143   11330.52 662.6929

stepwise = lm(y ~ position + country + hiprap + misc + pop +  south + northeast + midwest + west + paci
fic)
summary(stepwise)

##
## Call:
## lm(formula = y ~ position + country + hiprap + misc + pop + south +
##     northeast + midwest + west + pacific)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.194  -5.668  -0.119   4.605  37.889
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) 16.17882    5.07950   3.185  0.00179 **
## position    -0.21291    0.02665  -7.989 4.66e-13 ***
## country     10.35159    5.12907   2.018  0.04549 *
## hiprap      -0.64169    4.88568  -0.131  0.89570
## misc         7.34670    5.40161   1.360  0.17600
## pop          5.14472    4.90583   1.049  0.29614
## south       -0.48139    2.47353  -0.195  0.84598
## northeast   -1.33870    2.54849  -0.525  0.60022
## midwest      2.57441    3.46341   0.743  0.45854
## west         4.01951    2.99522   1.342  0.18179
## pacific     30.88938    9.23701   3.344  0.00106 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.976 on 139 degrees of freedom
## Multiple R-squared:  0.4732, Adjusted R-squared:  0.4353
## F-statistic: 12.49 on 10 and 139 DF,  p-value: 2.691e-15

anova(stepwise)

## Analysis of Variance Table
##
## Response: y
##             Df  Sum Sq Mean Sq F value    Pr(>F)
## position     1  6000.7  6000.7 74.4737 1.267e-14 ***
## country      1  1236.1  1236.1 15.3407 0.0001402 ***
## hiprap       1  1253.8  1253.8 15.5602 0.0001263 ***
## misc         1   152.6   152.6  1.8934 0.1710365
## pop          1    78.2    78.2  0.9703 0.3263293
## south        1    72.8    72.8  0.9041 0.3433409
## northeast    1   289.3   289.3  3.5900 0.0602065 .
## midwest      1     0.5     0.5  0.0062 0.9371388
## west         1    76.5    76.5  0.9488 0.3317059
## pacific      1   901.1   901.1 11.1829 0.0010613 **
## Residuals  139 11199.8    80.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

#b. Apply the backward elimination and obtain the solution.
step2<-stepAIC(entiremodel, direction="backward")

## Start:  AIC=676.31
## y ~ position + duration + feature + lifetime + numartists + singleoralbum +
##     country + hiprap + misc + pop + south + northeast + midwest +
##     west + pacific + tempo
##
##               Df Sum of Sq   RSS    AIC
## - hiprap       1       0.0 10859 674.31
## - tempo        1       0.1 10859 674.31
## - south        1       0.8 10860 674.32
## - duration     1      14.4 10873 674.51
## - numartists   1      18.9 10878 674.57
```

```
## - northeast      1      31.6 10890 674.75
## - midwest        1      53.8 10912 675.05
## - feature        1      63.8 10922 675.19
## - lifetime       1      94.9 10954 675.62
## - pop            1     112.2 10971 675.86
## - singleoralbum  1     129.9 10989 676.10
## <none>                       10859 676.31
## - west           1     147.4 11006 676.34
## - misc           1     234.9 11094 677.52
## - country        1     333.0 11192 678.84
## - pacific        1     858.9 11718 685.73
## - position       1    4843.8 15702 729.64
##
## Step:  AIC=674.31
## y ~ position + duration + feature + lifetime + numartists + singleoralbum +
##     country + misc + pop + south + northeast + midwest + west +
##     pacific + tempo
##
##                Df Sum of Sq    RSS    AIC
## - tempo         1       0.1 10859 672.31
## - south         1       0.8 10860 672.32
## - duration      1      14.6 10873 672.52
## - numartists    1      18.9 10878 672.57
## - northeast     1      31.6 10890 672.75
## - midwest       1      54.8 10914 673.07
## - feature       1      64.0 10923 673.20
## - lifetime      1      95.0 10954 673.62
## - singleoralbum 1     131.5 10990 674.12
## <none>                       10859 674.31
## - west          1     148.5 11007 674.35
## - pop           1     629.7 11488 680.77
## - misc          1     630.8 11490 680.78
## - pacific       1     858.9 11718 683.73
## - country       1    1770.7 12629 694.97
## - position      1    4902.7 15762 728.20
##
## Step:  AIC=672.31
## y ~ position + duration + feature + lifetime + numartists + singleoralbum +
##     country + misc + pop + south + northeast + midwest + west +
##     pacific
##
##                Df Sum of Sq    RSS    AIC
## - south         1       0.9 10860 670.33
## - duration      1      14.6 10873 670.52
## - numartists    1      18.9 10878 670.58
## - northeast     1      33.9 10893 670.78
## - midwest       1      56.5 10915 671.09
## - feature       1      64.0 10923 671.20
## - lifetime      1     101.5 10960 671.71
## - singleoralbum 1     132.8 10992 672.14
## <none>                       10859 672.31
```

```
## - west          1     148.9 11008 672.36
## - pop           1     634.8 11494 678.84
## - misc          1     663.3 11522 679.21
## - pacific       1     861.8 11721 681.77
## - country       1    1929.1 12788 694.84
## - position      1    4929.3 15788 726.46
##
## Step:  AIC=670.33
## y ~ position + duration + feature + lifetime + numartists + singleoralbum +
##     country + misc + pop + northeast + midwest + west + pacific
##
##               Df Sum of Sq   RSS    AIC
## - duration      1      14.0 10874 668.52
## - numartists    1      20.0 10880 668.60
## - northeast     1      41.1 10901 668.89
## - feature       1      63.4 10923 669.20
## - midwest       1      92.3 10952 669.60
## - lifetime      1     106.9 10967 669.80
## - singleoralbum 1     132.6 10992 670.15
## <none>                      10860 670.33
## - west          1     247.2 11107 671.70
## - pop           1     664.0 11524 677.23
## - misc          1     679.7 11540 677.43
## - pacific       1     880.4 11740 680.02
## - country       1    1963.3 12823 693.25
## - position      1    5531.6 16391 730.08
##
## Step:  AIC=668.52
## y ~ position + feature + lifetime + numartists + singleoralbum +
##     country + misc + pop + northeast + midwest + west + pacific
##
##               Df Sum of Sq   RSS    AIC
## - numartists    1      21.0 10895 666.81
## - northeast     1      46.6 10920 667.16
## - feature       1      63.5 10937 667.39
## - lifetime      1      95.4 10969 667.83
## - midwest       1      96.5 10970 667.84
## - singleoralbum 1     143.5 11017 668.49
## <none>                      10874 668.52
## - west          1     245.4 11119 669.87
## - pop           1     659.6 11533 675.35
## - misc          1     669.3 11543 675.48
## - pacific       1     885.9 11760 678.27
## - country       1    1949.9 12824 691.26
## - position      1    5563.6 16437 728.50
##
## Step:  AIC=666.81
## y ~ position + feature + lifetime + singleoralbum + country +
##     misc + pop + northeast + midwest + west + pacific
##
##               Df Sum of Sq   RSS    AIC
```

```
## - northeast      1      35.3 10930 665.30
## - lifetime       1      79.7 10974 665.90
## - feature        1      90.6 10985 666.05
## - midwest        1      96.2 10991 666.13
## - singleoralbum  1     134.9 11030 666.66
## <none>                      10895 666.81
## - west           1     224.9 11120 667.88
## - pop            1     644.2 11539 673.43
## - misc           1     672.8 11568 673.80
## - pacific        1     906.2 11801 676.79
## - country        1    1952.4 12847 689.54
## - position       1    5646.4 16541 727.45
##
## Step:  AIC=665.3
## y ~ position + feature + lifetime + singleoralbum + country +
##     misc + pop + midwest + west + pacific
##
##                Df Sum of Sq   RSS    AIC
## - lifetime       1      71.8 11002 664.28
## - feature        1      91.6 11022 664.55
## - midwest        1     118.9 11049 664.92
## - singleoralbum  1     126.7 11057 665.02
## <none>                      10930 665.30
## - west           1     291.8 11222 667.25
## - pop            1     609.2 11539 671.43
## - misc           1     648.6 11579 671.94
## - pacific        1     955.8 11886 675.87
## - country        1    2107.5 13038 689.74
## - position       1    5618.5 16549 725.51
##
## Step:  AIC=664.28
## y ~ position + feature + singleoralbum + country + misc + pop +
##     midwest + west + pacific
##
##                Df Sum of Sq   RSS    AIC
## - feature        1      91.3 11093 663.52
## - midwest        1     100.7 11102 663.64
## - singleoralbum  1     145.7 11148 664.25
## <none>                      11002 664.28
## - west           1     313.8 11316 666.50
## - pop            1     576.3 11578 669.94
## - misc           1     580.1 11582 669.98
## - pacific        1     964.1 11966 674.88
## - country        1    2065.8 13068 688.09
## - position       1    5555.7 16558 723.59
##
## Step:  AIC=663.52
## y ~ position + singleoralbum + country + misc + pop + midwest +
##     west + pacific
##
##                Df Sum of Sq   RSS    AIC
```

```
## - midwest        1     117.4 11210 663.10
## - singleoralbum  1     131.8 11225 663.29
## <none>                       11093 663.52
## - west           1     331.7 11425 665.94
## - misc           1     629.7 11723 669.80
## - pop            1     705.1 11798 670.76
## - pacific        1     970.4 12064 674.10
## - country        1    2272.7 13366 689.47
## - position       1    5484.0 16577 721.77
##
## Step:  AIC=663.1
## y ~ position + singleoralbum + country + misc + pop + west +
##     pacific
##
##                Df Sum of Sq    RSS    AIC
## - singleoralbum  1     120.0 11330 662.69
## <none>                       11210 663.10
## - west           1     290.7 11501 664.94
## - misc           1     621.5 11832 669.19
## - pop            1     650.3 11861 669.55
## - pacific        1     962.9 12173 673.46
## - country        1    2177.1 13388 687.72
## - position       1    5421.2 16632 720.26
##
## Step:  AIC=662.69
## y ~ position + country + misc + pop + west + pacific
##
##            Df Sum of Sq    RSS    AIC
## <none>                   11330 662.69
## - west      1     281.5 11612 664.37
## - misc      1     589.9 11920 668.31
## - pop       1     728.6 12059 670.04
## - pacific   1     961.5 12292 672.91
## - country   1    2247.9 13578 687.84
## - position  1    5696.3 17027 721.79

step2$anova # display results

## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## y ~ position + duration + feature + lifetime + numartists + singleoralbum +
##     country + hiprap + misc + pop + south + northeast + midwest +
##     west + pacific + tempo
##
## Final Model:
## y ~ position + country + misc + pop + west + pacific
##
##
##              Step Df     Deviance Resid. Df Resid. Dev      AIC
```

```
## 1                                133   10858.73 676.3134
## 2        - hiprap  1 5.121132e-04       134   10858.73 674.3134
## 3         - tempo  1 7.627170e-02       135   10858.81 672.3145
## 4         - south  1 8.998228e-01       136   10859.71 670.3269
## 5      - duration  1 1.396770e+01       137   10873.67 668.5197
## 6    - numartists  1 2.104988e+01       138   10894.72 666.8098
## 7     - northeast  1 3.531772e+01       139   10930.04 665.2953
## 8      - lifetime  1 7.175365e+01       140   11001.80 664.2768
## 9       - feature  1 9.134956e+01       141   11093.15 663.5171
## 10      - midwest  1 1.173522e+02       142   11210.50 663.0956
## 11 - singleoralbum  1 1.200187e+02       143   11330.52 662.6929

backwards = lm(y ~ position + country + hiprap + misc + pop +  south + northeast + midwest +
west + pacific)
summary(backwards)

##
## Call:
## lm(formula = y ~ position + country + hiprap + misc + pop + south +
##     northeast + midwest + west + pacific)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.194  -5.668  -0.119   4.605  37.889
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 16.17882    5.07950   3.185  0.00179 **
## position    -0.21291    0.02665  -7.989 4.66e-13 ***
## country     10.35159    5.12907   2.018  0.04549 *
## hiprap      -0.64169    4.88568  -0.131  0.89570
## misc         7.34670    5.40161   1.360  0.17600
## pop          5.14472    4.90583   1.049  0.29614
## south       -0.48139    2.47353  -0.195  0.84598
## northeast   -1.33870    2.54849  -0.525  0.60022
## midwest      2.57441    3.46341   0.743  0.45854
## west         4.01951    2.99522   1.342  0.18179
## pacific     30.88938    9.23701   3.344  0.00106 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.976 on 139 degrees of freedom
## Multiple R-squared:  0.4732, Adjusted R-squared:  0.4353
## F-statistic: 12.49 on 10 and 139 DF,  p-value: 2.691e-15

anova(backwards)

## Analysis of Variance Table
##
## Response: y
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## position    1  6000.7  6000.7 74.4737 1.267e-14 ***
```

```
## country     1  1236.1  1236.1 15.3407 0.0001402 ***
## hiprap      1  1253.8  1253.8 15.5602 0.0001263 ***
## misc        1   152.6   152.6  1.8934 0.1710365
## pop         1    78.2    78.2  0.9703 0.3263293
## south       1    72.8    72.8  0.9041 0.3433409
## northeast   1   289.3   289.3  3.5900 0.0602065 .
## midwest     1     0.5     0.5  0.0062 0.9371388
## west        1    76.5    76.5  0.9488 0.3317059
## pacific     1   901.1   901.1 11.1829 0.0010613 **
## Residuals 139 11199.8    80.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#c. Applythe best subsets regression and select a "best" model from the listof variables of this model.
# All Subsets Regression
library(leaps)
leaps<-regsubsets(y~position+duration+feature+lifetime+numartists+singleoralbum+country+hiprap+mi
sc+pop+south+northeast+midwest+west+pacific, nbest=2,data = testingdata,nvmax=10)# view results
summary(leaps)
```

```
## Subset selection object
## Call: regsubsets.formula(y ~ position + duration + feature + lifetime +
##     numartists + singleoralbum + country + hiprap + misc + pop +
##     south + northeast + midwest + west + pacific, nbest = 2,
##     data = testingdata, nvmax = 10)
## 15 Variables  (and intercept)
##             Forced in Forced out
## position        FALSE     FALSE
## duration        FALSE     FALSE
## feature         FALSE     FALSE
## lifetime        FALSE     FALSE
## numartists      FALSE     FALSE
## singleoralbum   FALSE     FALSE
## country         FALSE     FALSE
## hiprap          FALSE     FALSE
## misc            FALSE     FALSE
## pop             FALSE     FALSE
## south           FALSE     FALSE
## northeast       FALSE     FALSE
## midwest         FALSE     FALSE
## west            FALSE     FALSE
## pacific         FALSE     FALSE
## 2 subsets of each size up to 10
## Selection Algorithm: exhaustive
##          position duration feature lifetime numartists singleoralbum
## 1  ( 1 ) "*"      " "      " "     " "      " "         " "
## 1  ( 2 ) " "      " "      " "     " "      " "         " "
## 2  ( 1 ) "*"      " "      " "     " "      " "         " "
## 2  ( 2 ) "*"      " "      " "     " "      " "         " "
## 3  ( 1 ) "*"      " "      " "     " "      " "         " "
## 3  ( 2 ) "*"      " "      " "     " "      " "         " "
```

```
## 4  ( 1 ) "*"    " "    " "    " "    " "      " "
## 4  ( 2 ) "*"    " "    " "    " "    " "      " "
## 5  ( 1 ) "*"    " "    " "    " "    " "      " "
## 5  ( 2 ) "*"    " "    " "    " "    " "      " "
## 6  ( 1 ) "*"    " "    " "    " "    " "      " "
## 6  ( 2 ) "*"    " "    "*"    " "    " "      " "
## 7  ( 1 ) "*"    " "    " "    " "    " "      "*"
## 7  ( 2 ) "*"    " "    " "    " "    " "      " "
## 8  ( 1 ) "*"    " "    " "    " "    " "      "*"
## 8  ( 2 ) "*"    " "    "*"    " "    " "      "*"
## 9  ( 1 ) "*"    " "    "*"    " "    " "      "*"
## 9  ( 2 ) "*"    " "    " "    "*"    " "      "*"
## 10  ( 1 ) "*"    " "    "*"    "*"    " "      "*"
## 10  ( 2 ) "*"    " "    "*"    " "    " "      "*"
##           country hiprap misc pop south northeast midwest west pacific
## 1  ( 1 ) " "    " "   " "  " " " " "  " "     " "    " "  " "
## 1  ( 2 ) " "    "*"   " "  " " " " "  " "     " "    " "  " "
## 2  ( 1 ) " "    "*"   " "  " " " " "  " "     " "    " "  " "
## 2  ( 2 ) "*"    " "   " "  " " " " "  " "     " "    " "  " "
## 3  ( 1 ) " "    "*"   " "  " " " " "  " "     " "    " "  "*"
## 3  ( 2 ) " "    "*"   " "  " " " " "  "*"     " "    " "  " "
## 4  ( 1 ) "*"    "*"   " "  " " " " "  " "     " "    " "  "*"
## 4  ( 2 ) " "    "*"   " "  "*" " " "  " "     " "    " "  "*"
## 5  ( 1 ) "*"    "*"   " "  " " " " "  " "     " "    "*"  "*"
## 5  ( 2 ) "*"    " "   "*"  "*" " " "  " "     " "    " "  "*"
## 6  ( 1 ) "*"    " "   "*"  "*" " " "  " "     " "    "*"  "*"
## 6  ( 2 ) "*"    "*"   " "  " " " " "  " "     " "    "*"  "*"
## 7  ( 1 ) "*"    " "   "*"  "*" " " "  " "     " "    "*"  "*"
## 7  ( 2 ) "*"    " "   "*"  "*" " " "  " "     "*"    "*"  "*"
## 8  ( 1 ) "*"    " "   "*"  "*" " " "  " "     "*"    "*"  "*"
## 8  ( 2 ) "*"    " "   "*"  "*" " " "  " "     " "    "*"  "*"
## 9  ( 1 ) "*"    " "   "*"  "*" " " "  " "     "*"    "*"  "*"
## 9  ( 2 ) "*"    " "   "*"  "*" " " "  " "     "*"    "*"  "*"
## 10  ( 1 ) "*"    " "   "*"  "*" " " "  " "     "*"    "*"  "*"
## 10  ( 2 ) "*"    " "   "*"  "*" " " "  "*"     "*"    "*"  "*"
```

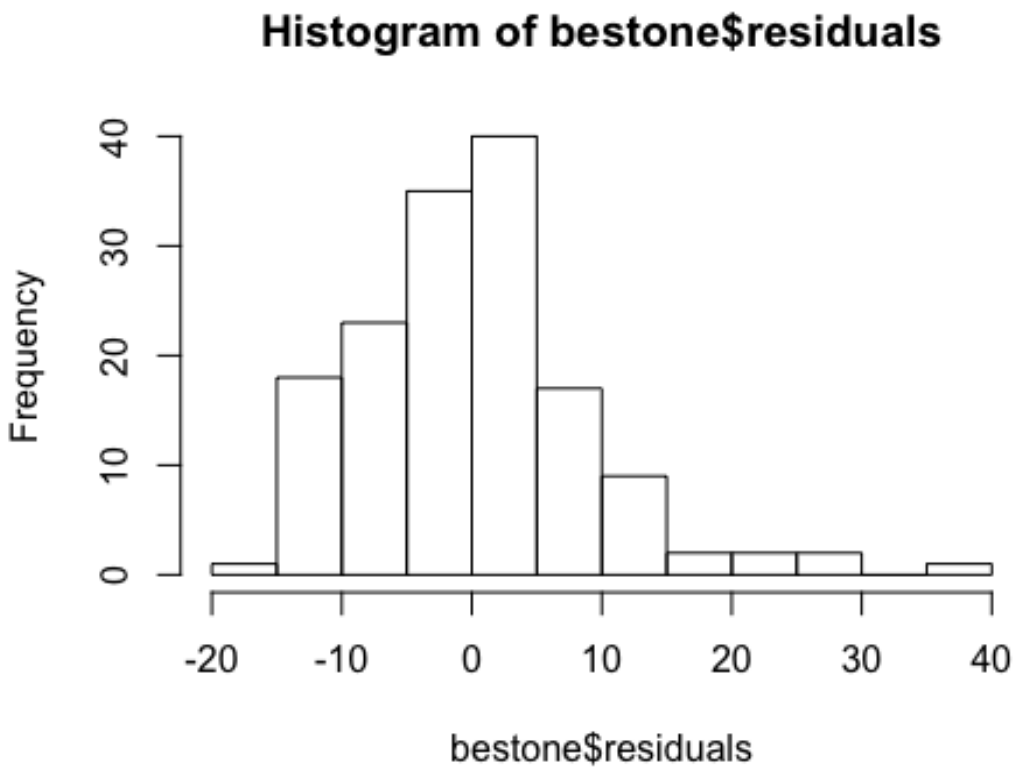
```
subsets1 = lm(y ~ position + feature + lifetime + singleoralbum + country + hiprap + misc + pop + south
+ northeast + midwest + west + pacific)
summary(subsets1)
```

```
##
## Call:
## lm(formula = y ~ position + feature + lifetime + singleoralbum +
##     country + hiprap + misc + pop + south + northeast + midwest +
##     west + pacific)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.585  -5.968  -0.376   4.362  37.561
##
## Coefficients:
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   13.03329    5.73568   2.272  0.02464 *
## position      -0.21519    0.02735  -7.867 1.01e-12 ***
## feature       -5.01447    4.69891  -1.067  0.28779
## lifetime      -0.08263    0.08553  -0.966  0.33568
## singleoralbum  3.20247    2.46573   1.299  0.19621
## country       10.73071    5.12517   2.094  0.03814 *
## hiprap         0.25364    4.90777   0.052  0.95886
## misc           9.41868    5.56085   1.694  0.09260 .
## pop            5.58377    4.90321   1.139  0.25679
## south         -0.28163    2.50230  -0.113  0.91055
## northeast     -1.52536    2.54878  -0.598  0.55052
## midwest        3.03534    3.50957   0.865  0.38863
## west           3.87064    2.99238   1.293  0.19803
## pacific       30.64356    9.21081   3.327  0.00113 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.95 on 136 degrees of freedom
## Multiple R-squared:  0.4876, Adjusted R-squared:  0.4387
## F-statistic: 9.956 on 13 and 136 DF,  p-value: 2.014e-14

anova(subsets1)

## Analysis of Variance Table
##
## Response: y
##               Df  Sum Sq Mean Sq F value    Pr(>F)
## position       1  6000.7  6000.7 74.9146 1.245e-14 ***
## feature        1   408.7   408.7  5.1024 0.0254836 *
## lifetime       1     0.4     0.4  0.0049 0.9443149
## singleoralbum  1   187.4   187.4  2.3398 0.1284250
## country        1  1111.0  1111.0 13.8696 0.0002861 ***
## hiprap         1   979.7   979.7 12.2305 0.0006356 ***
## misc           1   263.1   263.1  3.2843 0.0721493 .
## pop            1    84.9    84.9  1.0605 0.3049335
## south          1    54.4    54.4  0.6797 0.4111488
## northeast      1   317.5   317.5  3.9638 0.0484936 *
## midwest        1     4.6     4.6  0.0569 0.8117900
## west           1    68.7    68.7  0.8571 0.3561846
## pacific        1   886.6   886.6 11.0684 0.0011295 **
## Residuals    136 10893.6    80.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

subsets2 =  lm(y ~ position + feature + singleoralbum + country + hiprap + misc + pop + south + northeas
t + midwest + west + pacific)
summary(subsets2)

##
## Call:
## lm(formula = y ~ position + feature + singleoralbum + country +
```

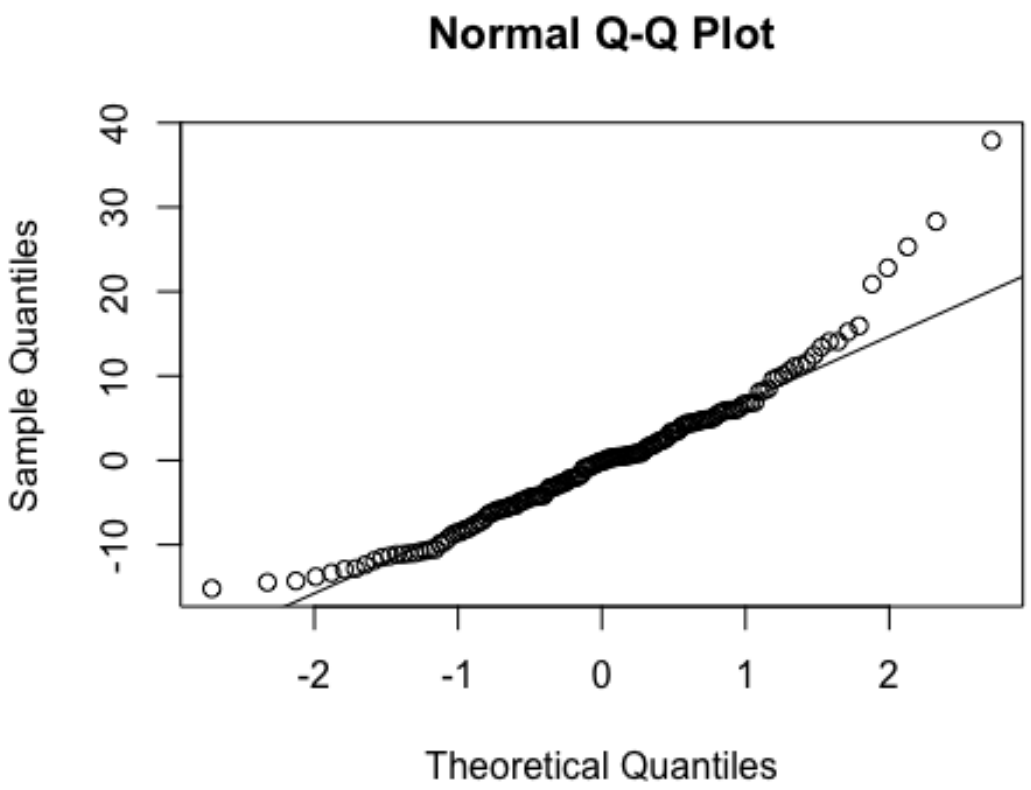
```
##     hiprap + misc + pop + south + northeast + midwest + west +
##     pacific)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.430  -5.733  -0.408   4.205  37.681
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   12.69178    5.72339   2.218  0.02823 *
## position      -0.20996    0.02681  -7.832 1.18e-12 ***
## feature       -5.03374    4.69773  -1.072  0.28582
## singleoralbum  3.41771    2.45505   1.392  0.16614
## country       10.77187    5.12374   2.102  0.03735 *
## hiprap         0.35382    4.90548   0.072  0.94261
## misc           8.24022    5.42411   1.519  0.13102
## pop            5.41988    4.89908   1.106  0.27053
## south         -0.67514    2.46833  -0.274  0.78487
## northeast     -1.60310    2.54689  -0.629  0.53011
## midwest        2.48804    3.46271   0.719  0.47366
## west           3.79907    2.99074   1.270  0.20614
## pacific       30.68060    9.20849   3.332  0.00111 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.948 on 137 degrees of freedom
## Multiple R-squared:  0.4841, Adjusted R-squared:  0.4389
## F-statistic: 10.71 on 12 and 137 DF,  p-value: 9.136e-15

anova(subsets2)

## Analysis of Variance Table
##
## Response: y
##               Df  Sum Sq Mean Sq F value    Pr(>F)
## position       1  6000.7  6000.7 74.9510 1.178e-14 ***
## feature        1   408.7   408.7  5.1049 0.0254369 *
## singleoralbum  1   186.8   186.8  2.3326 0.1289917
## country        1  1106.2  1106.2 13.8165 0.0002928 ***
## hiprap         1   972.0   972.0 12.1409 0.0006628 ***
## misc           1   194.6   194.6  2.4306 0.1212932
## pop            1    83.4    83.4  1.0423 0.3090934
## south          1    75.3    75.3  0.9411 0.3337101
## northeast      1   310.8   310.8  3.8817 0.0508323 .
## midwest        1     0.5     0.5  0.0066 0.9353946
## west           1    65.1    65.1  0.8126 0.3689349
## pacific        1   888.7   888.7 11.1007 0.0011095 **
## Residuals    137 10968.4    80.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
bestone = lm(y ~ position + country + hiprap + misc + pop +  south + northeast + midwest + west + pacific)
summary(bestone)

##
## Call:
## lm(formula = y ~ position + country + hiprap + misc + pop + south +
##     northeast + midwest + west + pacific)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.194  -5.668  -0.119   4.605  37.889
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 16.17882    5.07950   3.185  0.00179 **
## position    -0.21291    0.02665  -7.989 4.66e-13 ***
## country     10.35159    5.12907   2.018  0.04549 *
## hiprap      -0.64169    4.88568  -0.131  0.89570
## misc         7.34670    5.40161   1.360  0.17600
## pop          5.14472    4.90583   1.049  0.29614
## south       -0.48139    2.47353  -0.195  0.84598
## northeast   -1.33870    2.54849  -0.525  0.60022
## midwest      2.57441    3.46341   0.743  0.45854
## west         4.01951    2.99522   1.342  0.18179
## pacific     30.88938    9.23701   3.344  0.00106 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.976 on 139 degrees of freedom
## Multiple R-squared:  0.4732, Adjusted R-squared:  0.4353
## F-statistic: 12.49 on 10 and 139 DF,  p-value: 2.691e-15

#Assumption 1: E[residuals] = 0
hist(bestone$residuals) #histogram
```

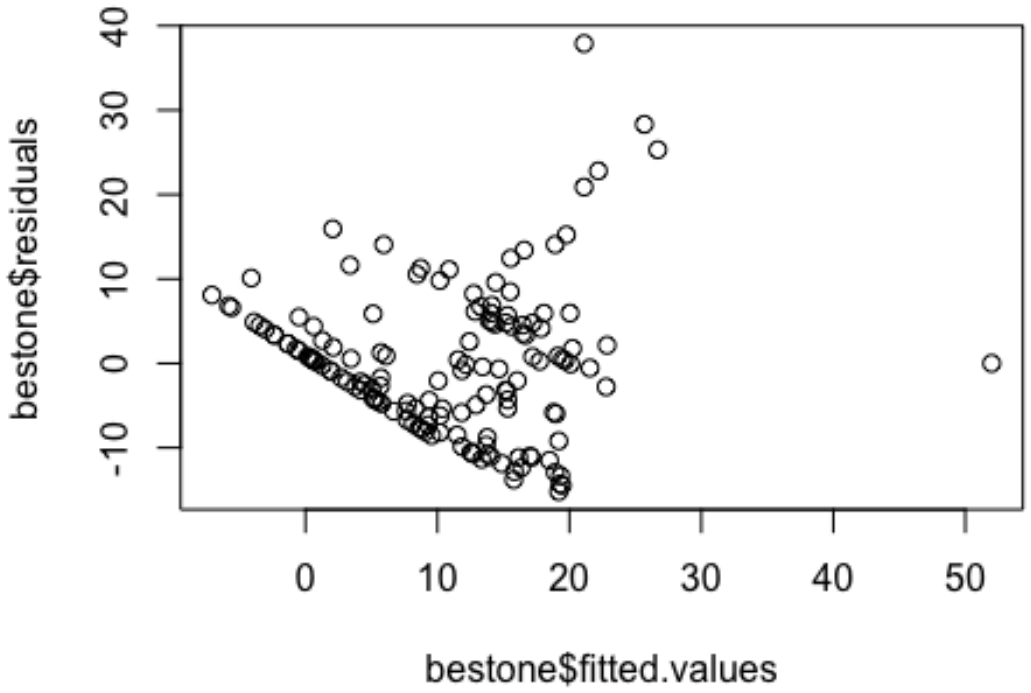


This code chunk looks into the assumptions of the "best model" from the screening methods (the above histogram, the mean of the residuals, QQplot and scatterplot).

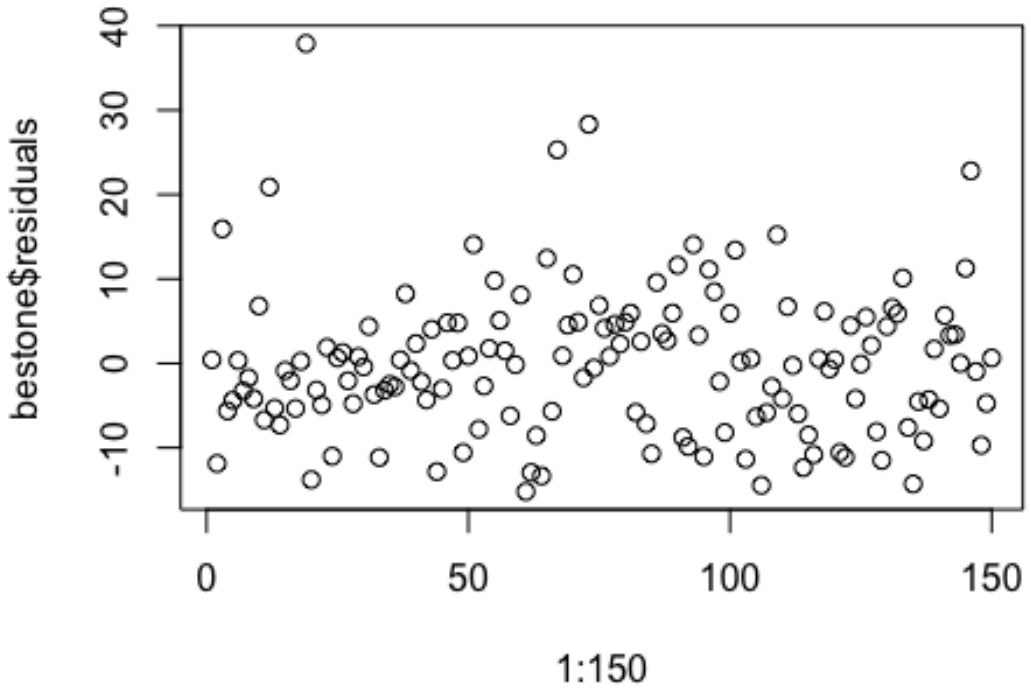
```
mean(bestone$residuals) #mean of residuals

## [1] -1.059696e-16

qqnorm(bestone$residuals)
qqline(bestone$residuals)
```




```
#Assumption 2: Normality
shapiro.test(bestone$residuals)

##
##  Shapiro-Wilk normality test
##
## data:  bestone$residuals
## W = 0.94458, p-value = 1.193e-05

#Assumption 3: Identical (constant) variance
plot(bestone$fitted.values,bestone$residuals)
```




```
plot(1:150,bestone$residuals)
```



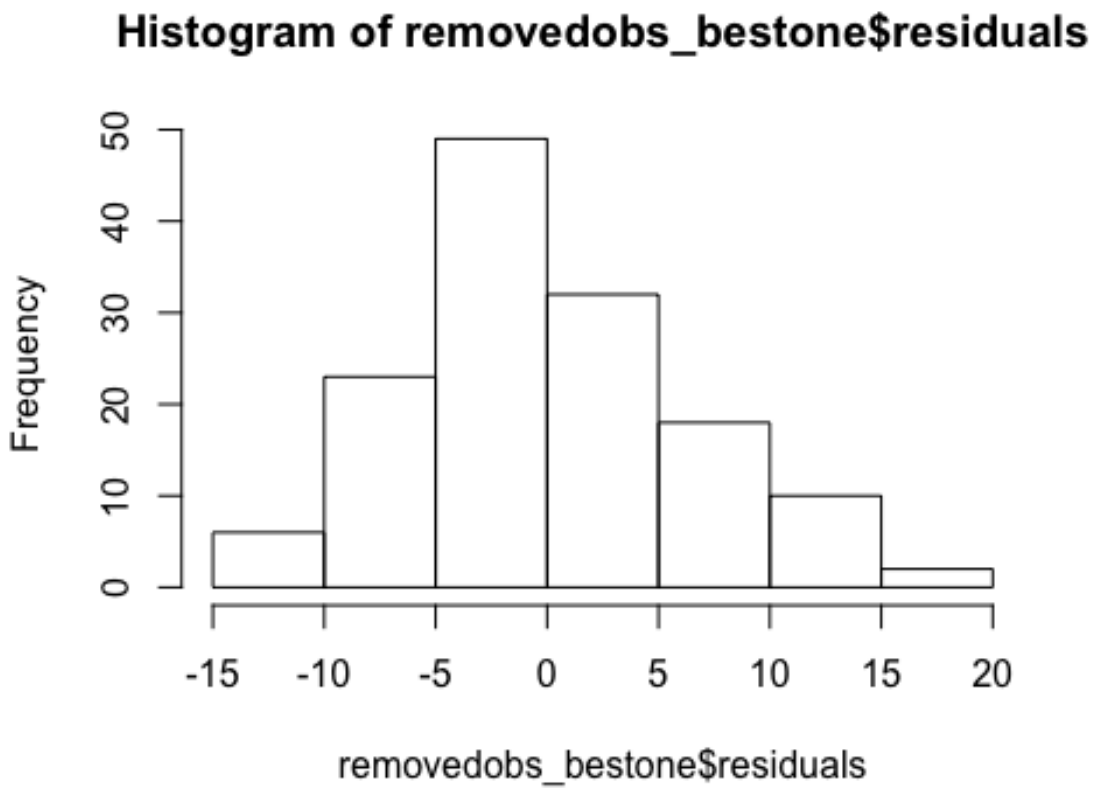
This code chunk redefines the data set after unusual observations and influential points were removed. A MLR model was then ran using the same variables from the "best model" from screening methods and with the addition of the change variable. It also again looks into the assumptions of the models' residuals (histogram, QQplot, scatterplot).

```
newdata <- read_excel("data_removed_unusualobs.xlsx")
attach(newdata)

## The following objects are masked _by_ .GlobalEnv:
##
##     change, feature, lifetime, tempo

## The following objects are masked from testingdata:
##
##     artist, change, D_c, D_h, D_m, D_mw, D_ne, D_p, D_pa, D_s,
##     D_w, duration_ms, feature, lifetime, number,
##     number_of_artists, peak_position, single_vs_album, song_name,
##     tempo, weeks_on_chart

## The following objects are masked from quantvars:
##
##     artist_origination, D_c, D_h, D_m, D_mw, D_ne, D_p, D_pa, D_s,
##     D_w, number
```

```
## The following objects are masked from data:
##
##     artist, artist_origination, change, duration_ms, feature,
##     lifetime, number, number_of_artists, peak_position,
##     single_vs_album, song_name, tempo, weeks_on_chart

newy=newdata$weeks_on_chart
newposition = newdata$peak_position
newchange = newdata$change
newtempo = newdata$tempo
newduration = newdata$duration_ms
newfeature = newdata$feature
newlifetime = newdata$lifetime
newnumartists = newdata$number_of_artists
newsingleoralbum = newdata$single_vs_album
#dummy variables for genre
newcountry = newdata$D_c
newhiprap = newdata$D_h
newmisc = newdata$D_m
newpop = newdata$D_p
#dummy variables for artist origination
newsouth = newdata$D_s
newnortheast = newdata$D_ne
newmidwest = newdata$D_mw
newwest = newdata$D_w
newpacific = newdata$D_pa

removedobs_bestone = lm(newy ~ newposition + newcountry + newhiprap + newmisc + newpop +  news
outh + newnortheast + newmidwest + newwest + newpacific + newchange)
summary(removedobs_bestone)

##
## Call:
## lm(formula = newy ~ newposition + newcountry + newhiprap + newmisc +
##     newpop + newsouth + newnortheast + newmidwest + newwest +
##     newpacific + newchange)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.575  -4.002  -1.078   3.626  15.920
##
## Coefficients: (2 not defined because of singularities)
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  16.67482    1.83533   9.085 1.47e-15 ***
## newposition  -0.18247    0.02141  -8.521 3.42e-14 ***
## newcountry    4.65725    1.82173   2.557  0.01172 *
## newhiprap    -4.12035    1.43170  -2.878  0.00468 **
## newmisc      -2.30305    2.33915  -0.985  0.32666
## newpop            NA         NA      NA       NA
## newsouth      2.84051    1.93021   1.472  0.14354
## newnortheast  3.78417    1.99816   1.894  0.06047 .
```

```
## newmidwest    5.85300    2.73235   2.142  0.03405 *
## newwest       6.88736    2.29721   2.998  0.00326 **
## newpacific         NA         NA      NA       NA
## newchange     0.17099    0.04079   4.192 5.07e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.608 on 130 degrees of freedom
## Multiple R-squared:  0.4824, Adjusted R-squared:  0.4465
## F-statistic: 13.46 on 9 and 130 DF,  p-value: 4.555e-15
```

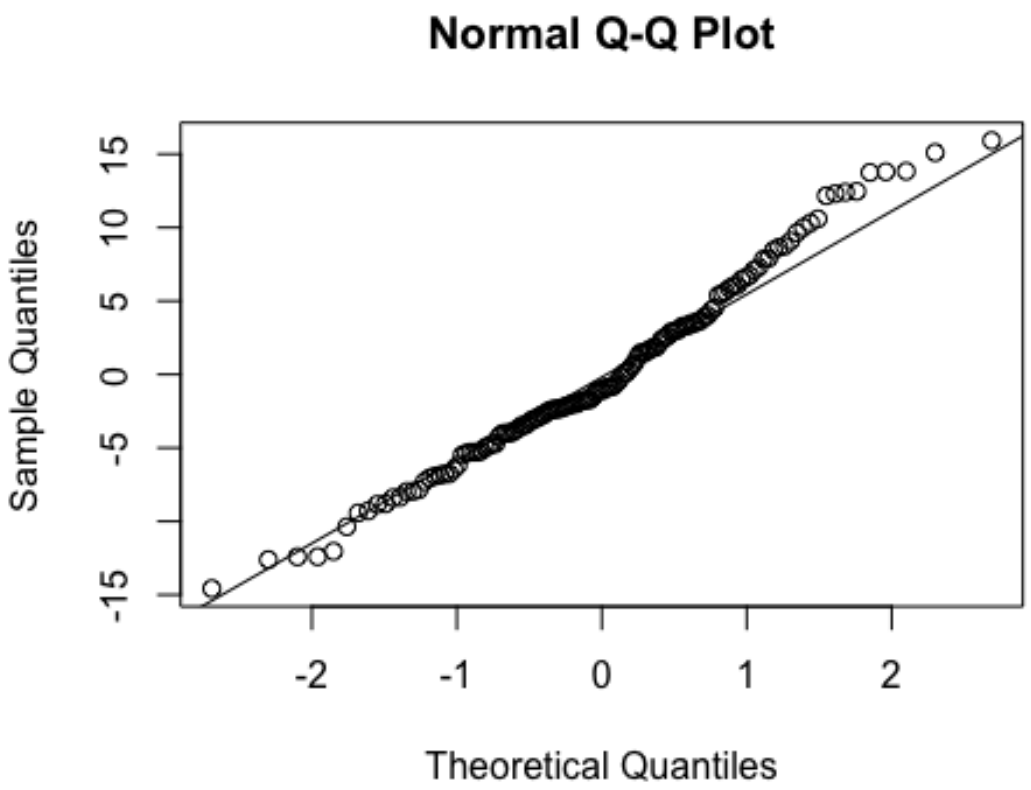
```
#Assumption 1: E[residuals] = 0
hist(removedobs_bestone$residuals) #histogram
```



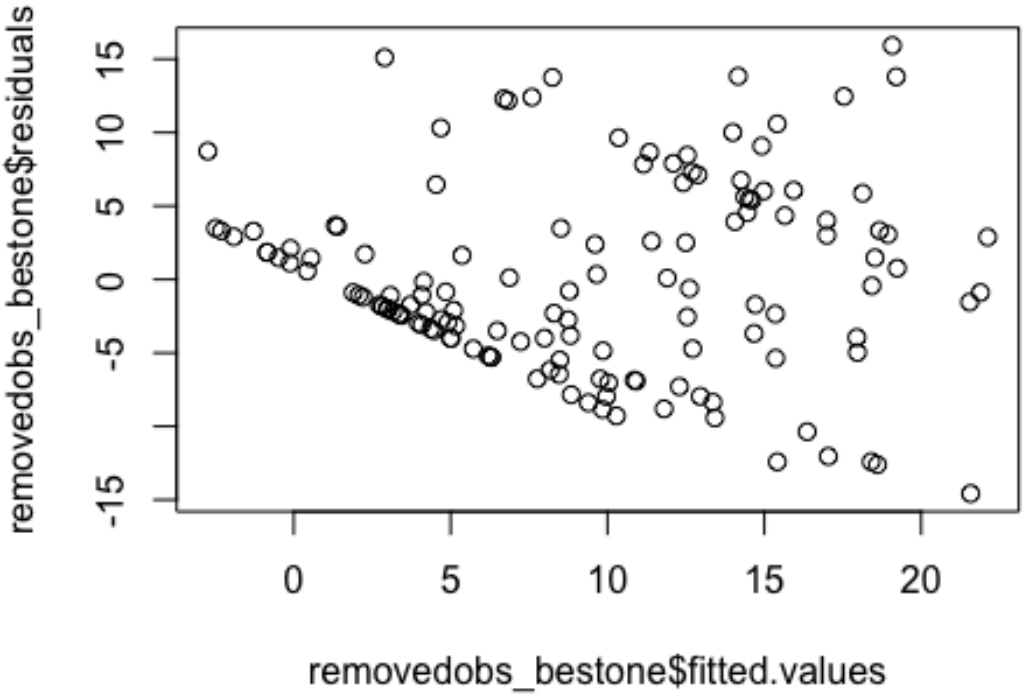

```
mean(removedobs_bestone$residuals) #mean of residuals
```

```
## [1] 9.068577e-19
```

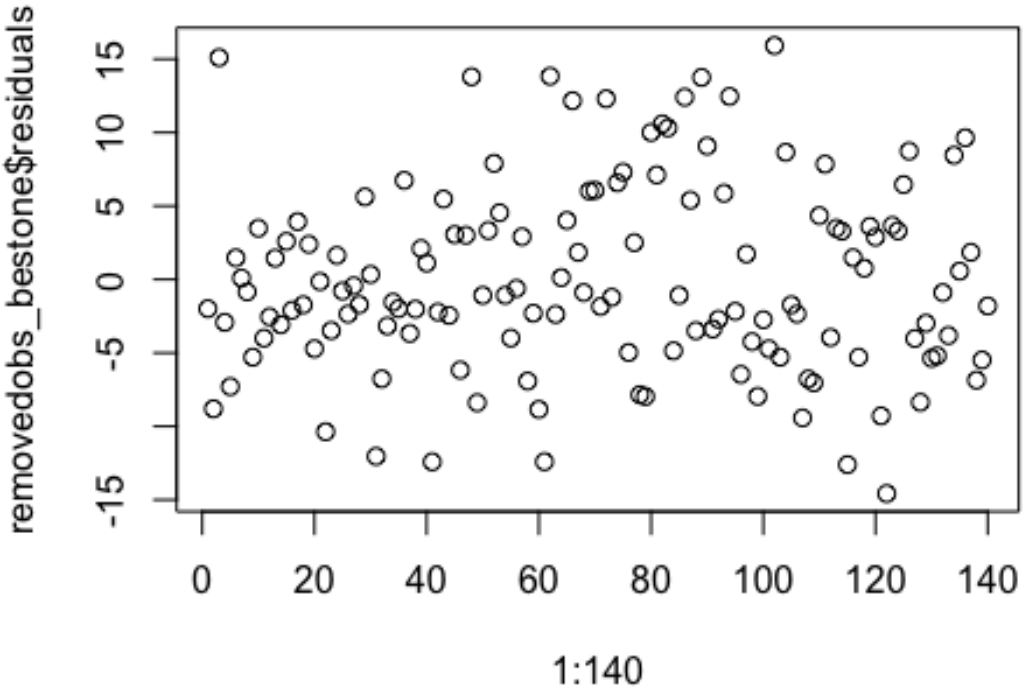
```
qqnorm(removedobs_bestone$residuals)
qqline(removedobs_bestone$residuals)
```




```
#Assumption 2: Normality
shapiro.test(removedobs_bestone$residuals)

##
##  Shapiro-Wilk normality test
##
## data:  removedobs_bestone$residuals
## W = 0.98362, p-value = 0.09273

#Assumption 3: Identical (constant) variance
plot(removedobs_bestone$fitted.values,removedobs_bestone$residuals)
```



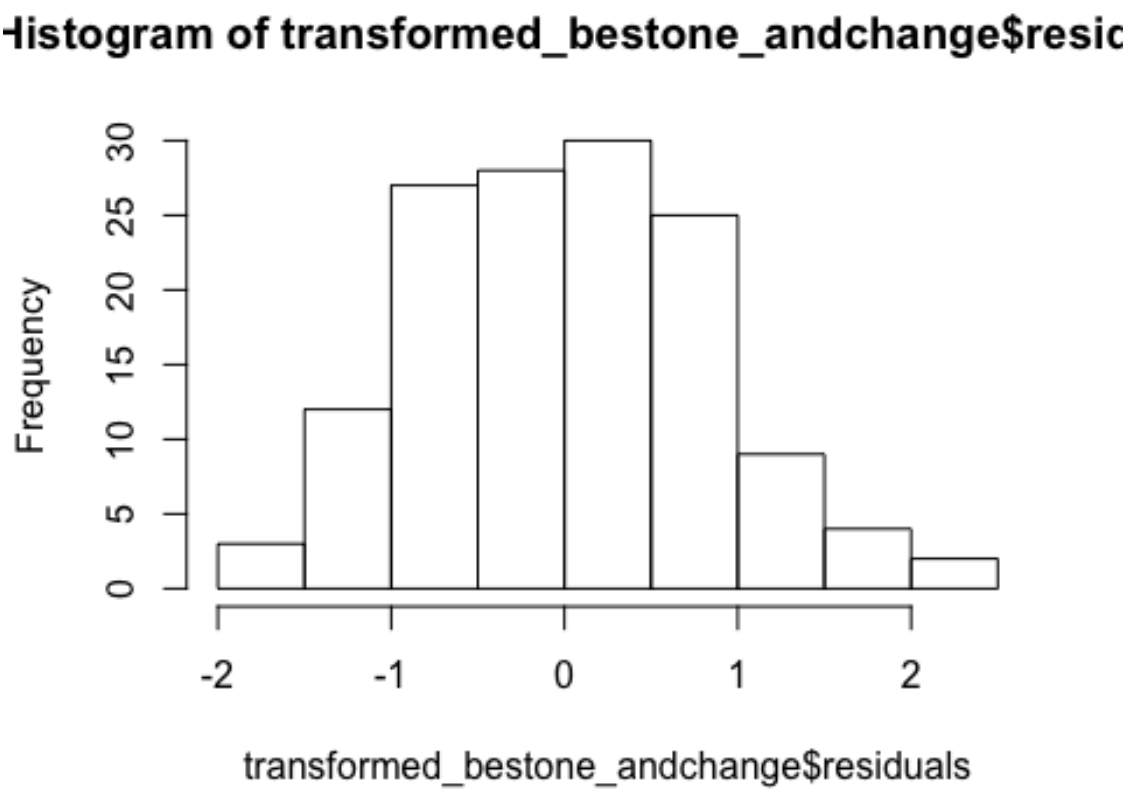

```
plot(1:140,removedobs_bestone$residuals)
```



This code chunk transforms the model with ln(y) and then runs a MLR model . It looks at the assumptions of the residuals through the same graphs, and ANOVA table.

```
transformed_bestone_andchange = lm(log(newy) ~ newposition + newcountry + newhiprap + newmisc +
newpop +  newsouth + newnortheast + newmidwest + newwest + newpacific +newchange)
summary(transformed_bestone_andchange)

##
## Call:
## lm(formula = log(newy) ~ newposition + newcountry + newhiprap +
##     newmisc + newpop + newsouth + newnortheast + newmidwest +
##     newwest + newpacific + newchange)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.65303 -0.57841 -0.00374  0.60934  2.28346
##
## Coefficients: (2 not defined because of singularities)
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.744816   0.234574  11.701  < 2e-16 ***
## newposition  -0.027745   0.002737 -10.138  < 2e-16 ***
## newcountry    0.763459   0.232835   3.279  0.00134 **
## newhiprap    -0.598959   0.182986  -3.273  0.00136 **
## newmisc      -0.103654   0.298967  -0.347  0.72937
```

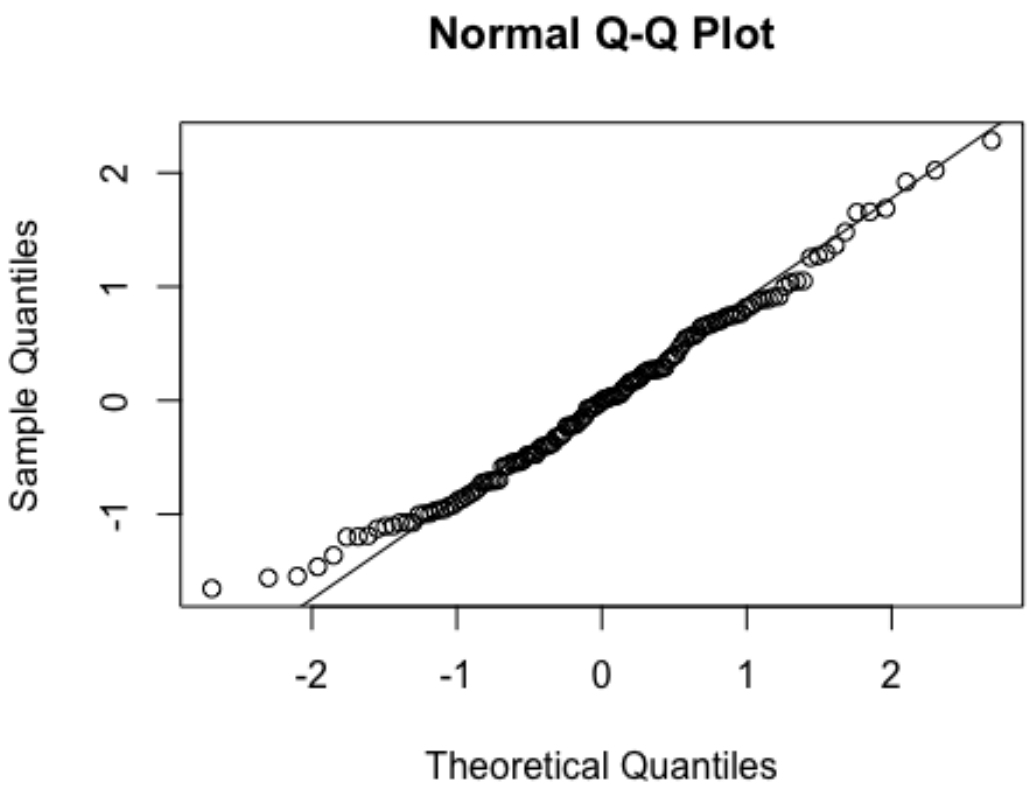
```
## newpop             NA        NA      NA       NA
## newsouth      0.284029   0.246700   1.151  0.25172
## newnortheast  0.354882   0.255385   1.390  0.16703
## newmidwest    0.950438   0.349222   2.722  0.00739 **
## newwest       0.955031   0.293607   3.253  0.00146 **
## newpacific         NA         NA      NA       NA
## newchange     0.016317   0.005213   3.130  0.00216 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8445 on 130 degrees of freedom
## Multiple R-squared:  0.5556, Adjusted R-squared:  0.5249
## F-statistic: 18.06 on 9 and 130 DF,  p-value: < 2.2e-16

anova(transformed_bestone_andchange)

## Analysis of Variance Table
##
## Response: log(newy)
##              Df Sum Sq Mean Sq F value    Pr(>F)
## newposition    1 60.554  60.554 84.9047 7.099e-16 ***
## newcountry     1 22.793  22.793 31.9589 9.529e-08 ***
## newhiprap      1 12.963  12.963 18.1757 3.842e-05 ***
## newmisc        1  0.283   0.283  0.3964  0.530072
## newsouth       1  1.800   1.800  2.5239  0.114564
## newnortheast   1  1.132   1.132  1.5870  0.210017
## newmidwest     1  1.591   1.591  2.2313  0.137665
## newwest        1  7.823   7.823 10.9682  0.001200 **
## newchange      1  6.988   6.988  9.7976  0.002158 **
## Residuals    130 92.716   0.713
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

#Assumption 1: E[residuals] = 0
hist(transformed_bestone_andchange$residuals) #histogram
```



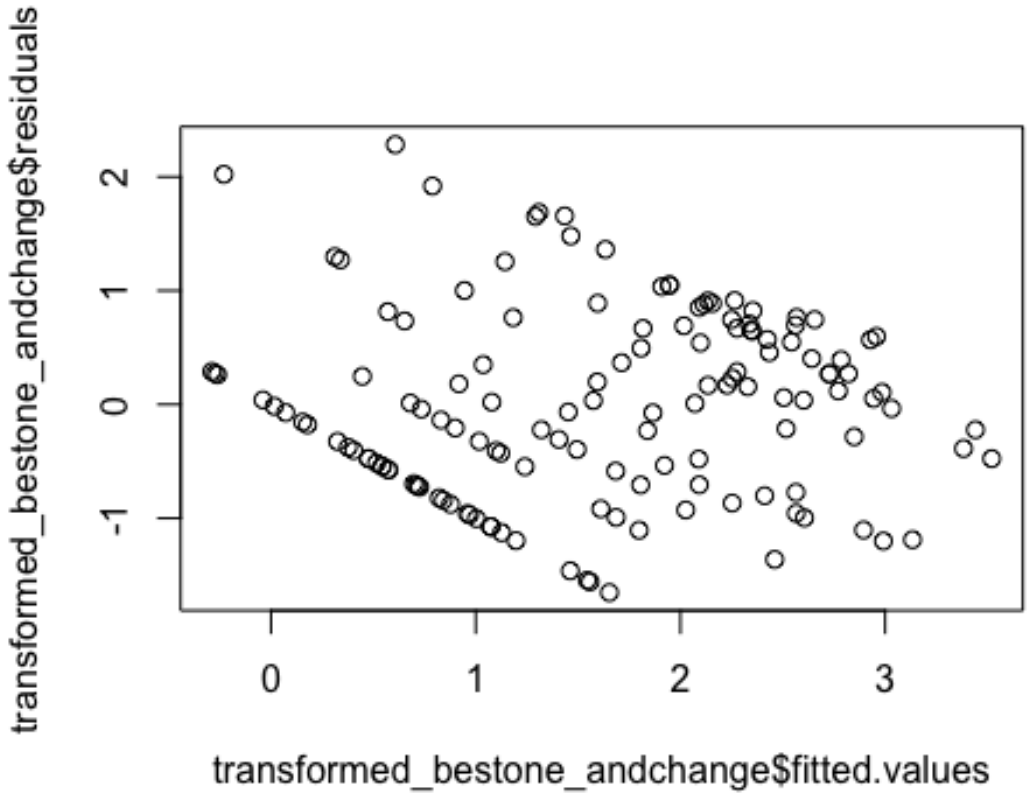

```
mean(transformed_bestone_andchange$residuals) #mean of residuals
```

```
## [1] 3.453765e-17
```

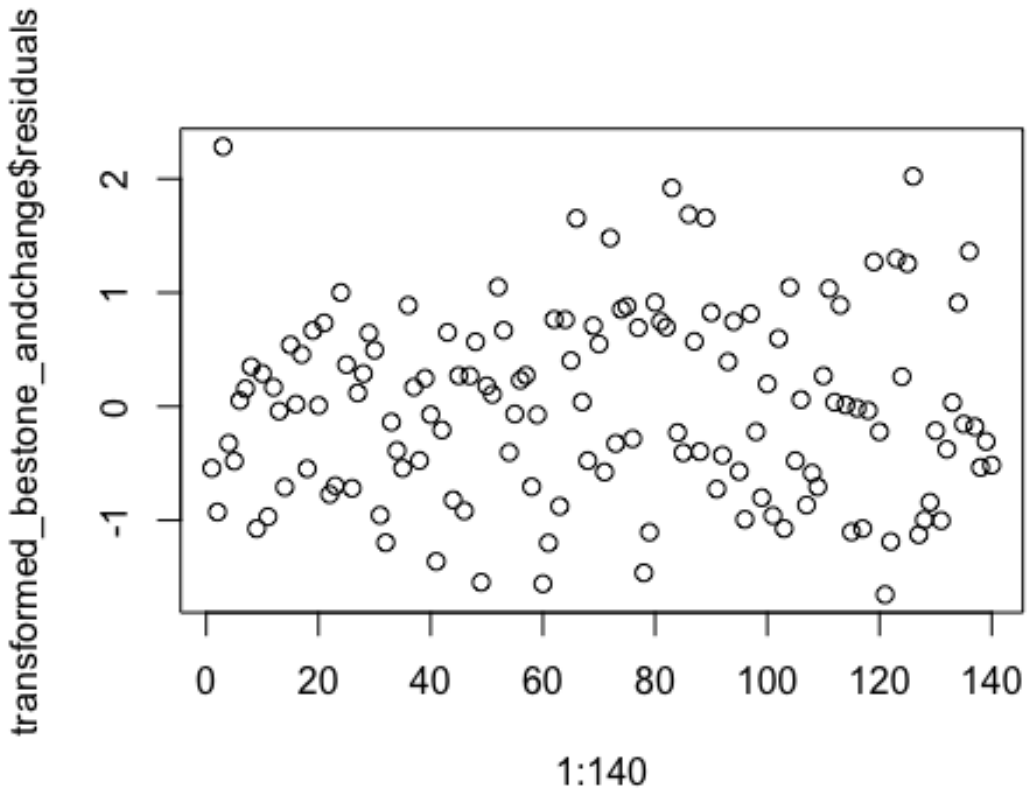
```
qqnorm(transformed_bestone_andchange$residuals)
qqline(transformed_bestone_andchange$residuals)
```




```
#Assumption 2: Normality
shapiro.test(transformed_bestone_andchange$residuals)

##
##  Shapiro-Wilk normality test
##
## data:  transformed_bestone_andchange$residuals
## W = 0.98645, p-value = 0.1853

#Assumption 3: Identical (constant) variance
plot(transformed_bestone_andchange$fitted.values,transformed_bestone_andchange$residuals)
```




```
plot(1:140,transformed_bestone_andchange$residuals)
```


transformed_bestone_andchange$residuals
-1
0
1
2
0
20
40
60
80
100
120
140
1:140

7. References

Forecasting, Time Series, and Regression:

Bowerman, Bruce L., et al. *Forecasting, Time Series, and Regression: an Applied Approach*. Thomson Brooks/Cole, 2005.

Spotify API: https://developer.spotify.com/documentation/web-api/