

4_Detailed_Report

Hyungjae Kim

2022-08-12

Dataset

Raw Dataset

Removed track_hist, url, artists_ids, track_id, chart_start, chart_end, and new_id.

	hitness	danceability	energy	key	loudness	mode	speechiness	acousticness
1:	1260	0.841	0.728	7	-3.370	1	0.0484	0.0847
2:	2560	0.336	0.627	7	-7.463	1	0.0384	0.1640
3:	1490	0.589	0.472	8	-8.749	1	0.0502	0.6140
4:	1140	0.754	0.424	2	-8.463	1	0.0363	0.6430
5:	1000	0.683	0.375	0	-13.056	1	0.0303	0.5790
6:	1070	0.240	0.598	7	-8.435	1	0.0369	0.7660

	instrumentalness	liveness	valence	tempo	duration_ms	time_signature
1:	0	0.1490	0.430	130.049	243837	4
2:	0	0.0708	0.350	150.273	241107	4
3:	0	0.5050	0.898	67.196	126267	4
4:	0	0.0652	0.806	119.705	130973	4
5:	0	0.0760	0.888	140.467	135533	4
6:	0	0.1170	0.776	201.629	151933	3

	popularity	sentiment
1:	17532.333	0.9938
2:	2772.143	0.9805
3:	410.000	0.9868
4:	281.000	0.8442
5:	1169.000	0.9966
6:	1144.000	0.9973

Classified Dataset

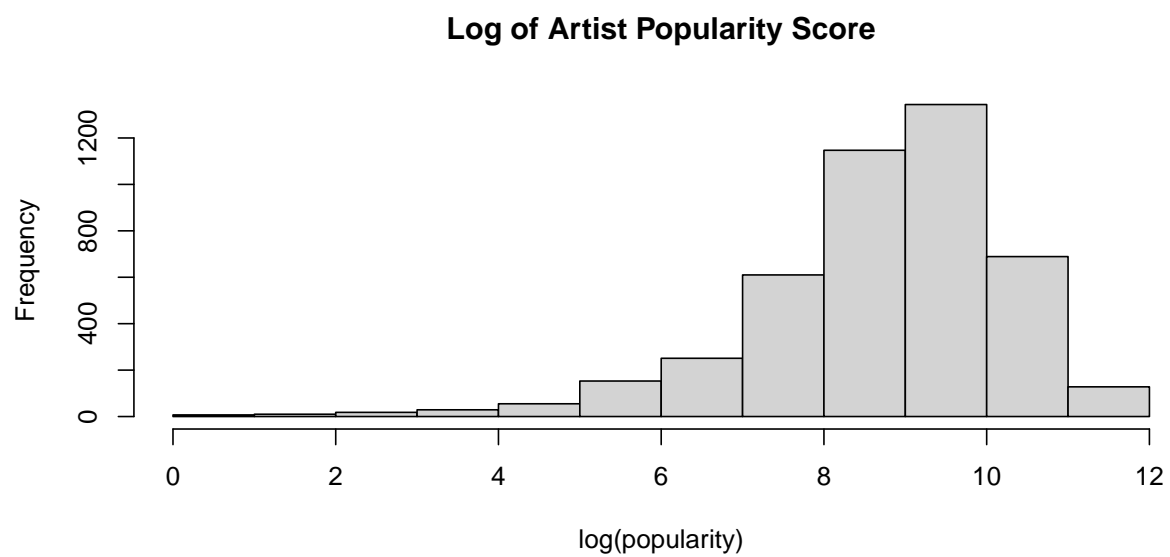
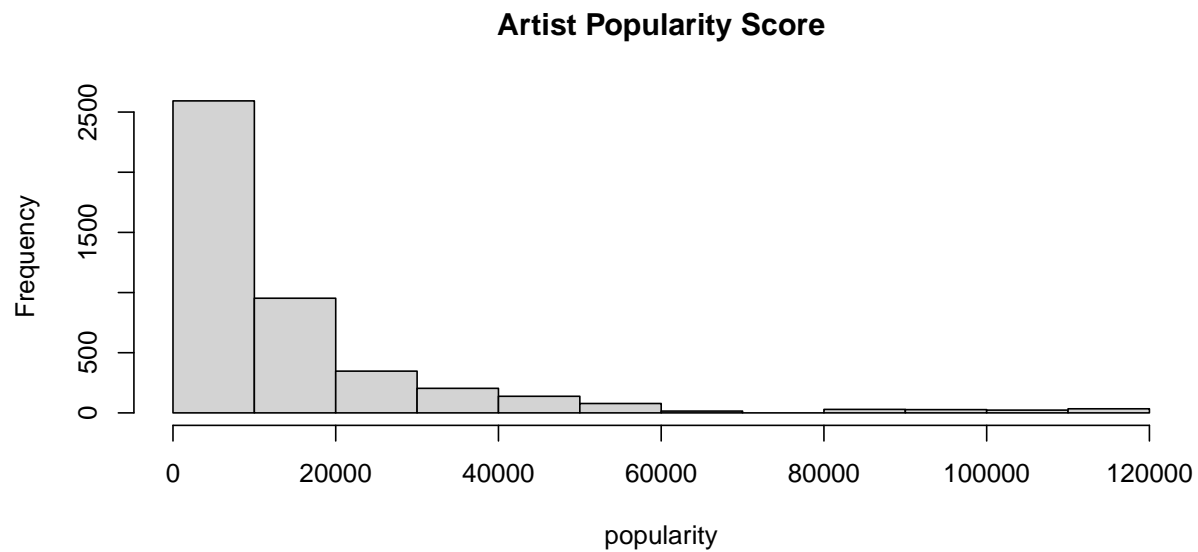
Classified songs as hit songs by using a **hitness** threshold of 250, converted duration from milliseconds to minutes, and log scaled popularity and tempo. This dataset was used for all the models.

	hitness	danceability	energy	key	loudness	mode	speechiness	acousticness
1:	TRUE	0.841	0.728	7	-3.370	1	0.0484	0.0847
2:	TRUE	0.336	0.627	7	-7.463	1	0.0384	0.1640
3:	TRUE	0.589	0.472	8	-8.749	1	0.0502	0.6140
4:	TRUE	0.754	0.424	2	-8.463	1	0.0363	0.6430
5:	TRUE	0.683	0.375	0	-13.056	1	0.0303	0.5790
6:	TRUE	0.240	0.598	7	-8.435	1	0.0369	0.7660

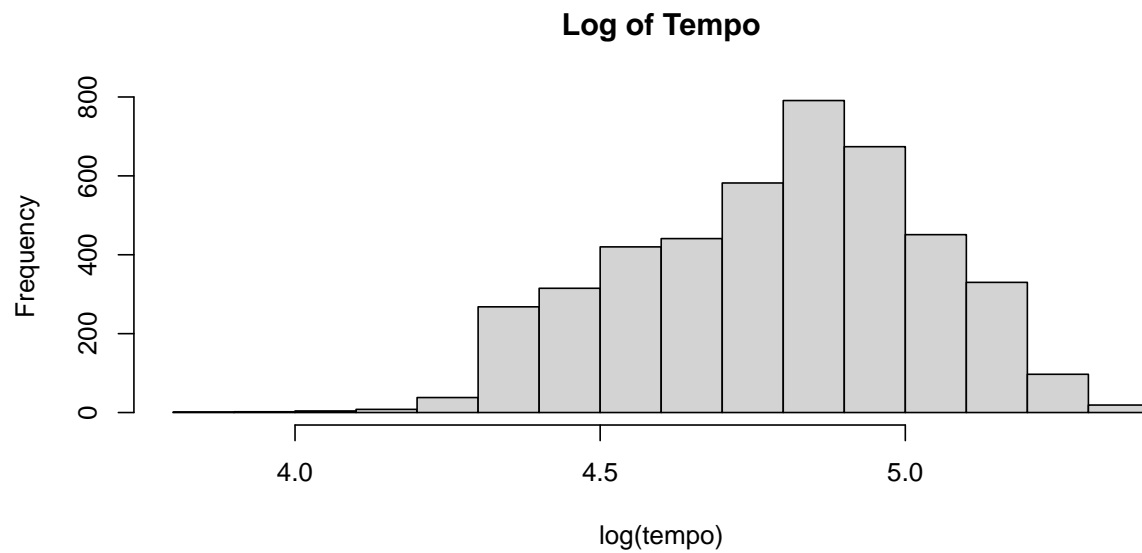
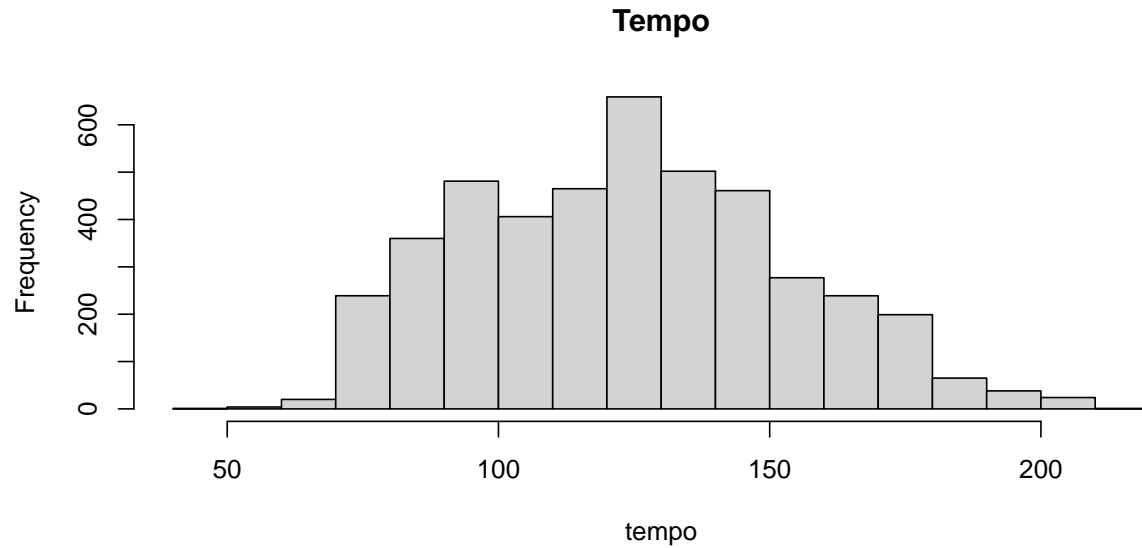
	instrumentalness	liveness	valence	time_signature	sentiment	duration_min
1:	0	0.1490	0.430	4	0.9938	4.063950
2:	0	0.0708	0.350	4	0.9805	4.018450
3:	0	0.5050	0.898	4	0.9868	2.104450
4:	0	0.0652	0.806	4	0.8442	2.182883
5:	0	0.0760	0.888	4	0.9966	2.258883
6:	0	0.1170	0.776	3	0.9973	2.532217

	log_tempo	log_popularity
1:	4.867911	9.771802
2:	5.012454	7.927376
3:	4.207614	6.016157
4:	4.785030	5.638355
5:	4.944973	7.063904
6:	5.306429	7.042286

Log Scaling Popularity



Log Scaling Tempo



Number of Rows / Feature Vectors

[1] 4441

Summary of Raw Data (Continuous Hitness Values and Unscaled)

hitness	danceability	energy	key
Min. : 20.0	Min. :0.0671	Min. :0.0326	Min. : 0.000
1st Qu.: 40.0	1st Qu.:0.5290	1st Qu.:0.5700	1st Qu.: 2.000
Median : 220.0	Median :0.6300	Median :0.7090	Median : 5.000

Mean : 419.5	Mean :0.6251	Mean :0.6822	Mean : 5.208
3rd Qu.: 480.0	3rd Qu.:0.7290	3rd Qu.:0.8180	3rd Qu.: 8.000
Max. :4250.0	Max. :0.9860	Max. :0.9960	Max. :11.000
loudness	mode	speechiness	acousticness
Min. : -29.224	Min. :0.0000	Min. :0.0225	Min. :0.0000033
1st Qu.: -7.054	1st Qu.:0.0000	1st Qu.:0.0368	1st Qu.:0.0183000
Median : -5.616	Median :1.0000	Median :0.0563	Median :0.0736000
Mean : -5.964	Mean :0.6715	Mean :0.1062	Mean :0.1718459
3rd Qu.: -4.403	3rd Qu.:1.0000	3rd Qu.:0.1280	3rd Qu.:0.2410000
Max. : 0.175	Max. :1.0000	Max. :0.9510	Max. :0.9930000
instrumentalness	liveness	valence	tempo
Min. :0.0000000	Min. :0.0193	Min. :0.0349	Min. : 48.72
1st Qu.:0.0000000	1st Qu.:0.0979	1st Qu.:0.3190	1st Qu.:100.01
Median :0.0000000	Median :0.1310	Median :0.4780	Median :124.08
Mean :0.0112172	Mean :0.1892	Mean :0.4861	Mean :123.97
3rd Qu.:0.0000131	3rd Qu.:0.2460	3rd Qu.:0.6510	3rd Qu.:143.89
Max. :0.9550000	Max. :0.9790	Max. :0.9760	Max. :213.74
duration_ms	time_signature	popularity	sentiment
Min. : 46253	Min. :1.000	Min. : 1	Min. : -1.0000
1st Qu.:197759	1st Qu.:4.000	1st Qu.: 2938	1st Qu.: -0.7613
Median :219840	Median :4.000	Median : 7716	Median : 0.9636
Mean :224905	Mean :3.973	Mean : 13929	Mean : 0.3771
3rd Qu.:245867	3rd Qu.:4.000	3rd Qu.: 16468	3rd Qu.: 0.9949
Max. :688453	Max. :5.000	Max. :115282	Max. : 1.0000

Summary of Classified and Cleaned Data (Factorized and Scaled)

hitness	danceability	energy	key
Mode :logical	Min. :0.0671	Min. :0.0326	Min. : 0.000
FALSE:2378	1st Qu.:0.5290	1st Qu.:0.5700	1st Qu.: 2.000
TRUE :2063	Median :0.6300	Median :0.7090	Median : 5.000
	Mean :0.6251	Mean :0.6822	Mean : 5.208
	3rd Qu.:0.7290	3rd Qu.:0.8180	3rd Qu.: 8.000
	Max. :0.9860	Max. :0.9960	Max. :11.000
loudness	mode	speechiness	acousticness
Min. : -29.224	Min. :0.0000	Min. :0.0225	Min. :0.0000033
1st Qu.: -7.054	1st Qu.:0.0000	1st Qu.:0.0368	1st Qu.:0.0183000
Median : -5.616	Median :1.0000	Median :0.0563	Median :0.0736000
Mean : -5.964	Mean :0.6715	Mean :0.1062	Mean :0.1718459
3rd Qu.: -4.403	3rd Qu.:1.0000	3rd Qu.:0.1280	3rd Qu.:0.2410000
Max. : 0.175	Max. :1.0000	Max. :0.9510	Max. :0.9930000
instrumentalness	liveness	valence	time_signature
Min. :0.0000000	Min. :0.0193	Min. :0.0349	Min. :1.000
1st Qu.:0.0000000	1st Qu.:0.0979	1st Qu.:0.3190	1st Qu.:4.000
Median :0.0000000	Median :0.1310	Median :0.4780	Median :4.000
Mean :0.0112172	Mean :0.1892	Mean :0.4861	Mean :3.973
3rd Qu.:0.0000131	3rd Qu.:0.2460	3rd Qu.:0.6510	3rd Qu.:4.000
Max. :0.9550000	Max. :0.9790	Max. :0.9760	Max. :5.000
sentiment	duration_min	log_tempo	log_popularity
Min. : -1.0000	Min. : 0.7709	Min. :3.886	Min. : 0.000
1st Qu.: -0.7613	1st Qu.: 3.2960	1st Qu.:4.605	1st Qu.: 7.985
Median : 0.9636	Median : 3.6640	Median :4.821	Median : 8.951
Mean : 0.3771	Mean : 3.7484	Mean :4.791	Mean : 8.708

3rd Qu.: 0.9949	3rd Qu.: 4.0978	3rd Qu.:4.969	3rd Qu.: 9.709
Max. : 1.0000	Max. :11.4742	Max. :5.365	Max. :11.655

Linear Model (OLS)

Accuracy

[1] 0.6147615

Summary (Coefficients, F-Statistics, P-Value)

Call:

```
lm(formula = hitness ~ ., data = dataset.5.train)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.7541	-0.4439	-0.2277	0.4886	0.9315

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-0.267899	0.401848	-0.667	0.5051	
danceability	0.255939	0.118198	2.165	0.0306	*
energy	-0.206569	0.150078	-1.376	0.1690	
key	0.001504	0.004107	0.366	0.7143	
loudness	0.014674	0.009391	1.563	0.1184	
mode	0.072420	0.032524	2.227	0.0262	*
speechiness	-0.594191	0.145641	-4.080	4.83e-05	***
acousticness	-0.159829	0.085221	-1.875	0.0610	.
instrumentalness	-0.008887	0.132692	-0.067	0.9466	
liveness	0.010336	0.106168	0.097	0.9225	
valence	0.261433	0.079235	3.299	0.0010	***
time_signature	0.054399	0.052688	1.032	0.3021	
sentiment	0.024539	0.018323	1.339	0.1808	
duration_min	0.006825	0.018677	0.365	0.7149	
log_tempo	0.005976	0.060501	0.099	0.9213	
log_popularity	0.047394	0.009625	4.924	9.77e-07	***

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

Residual standard error: 0.4844 on 1095 degrees of freedom

Multiple R-squared: 0.06718, Adjusted R-squared: 0.0544

F-statistic: 5.257 on 15 and 1095 DF, p-value: 2.906e-10

Anova Table

Anova Table (Type II tests)

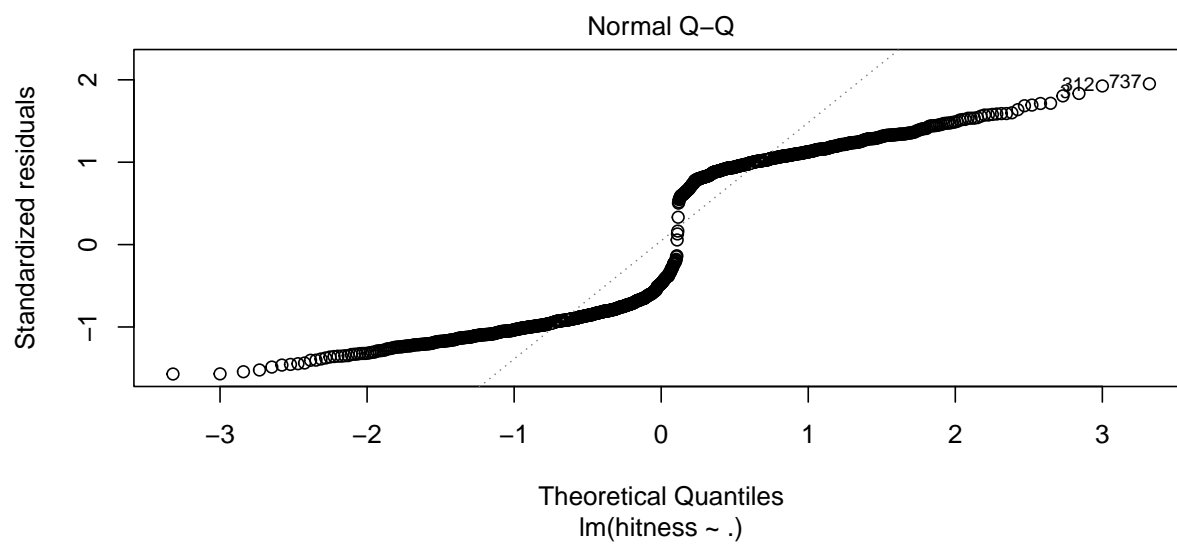
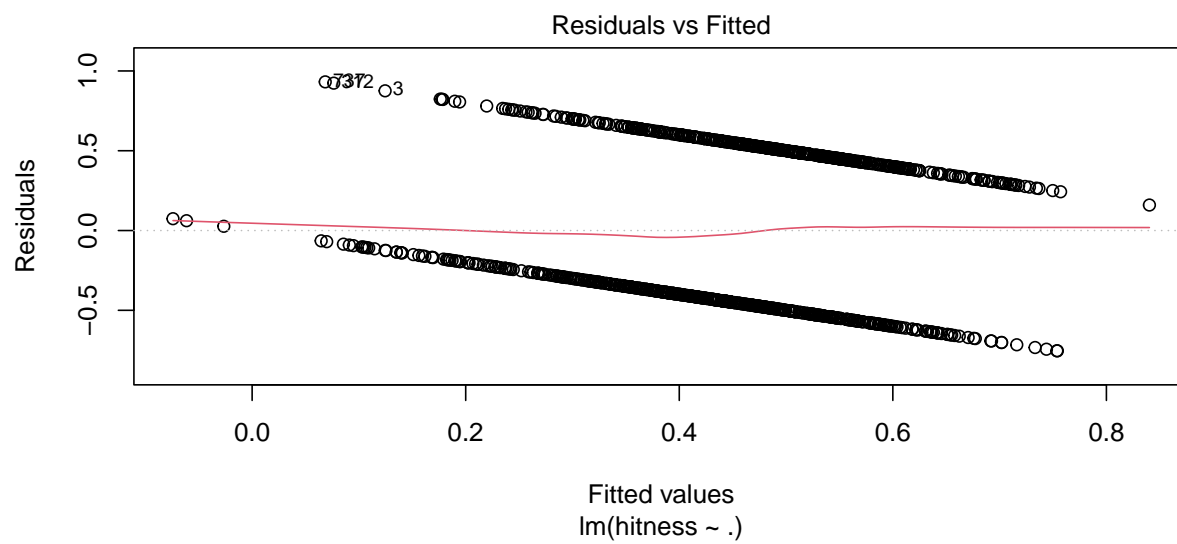
Response: hitness

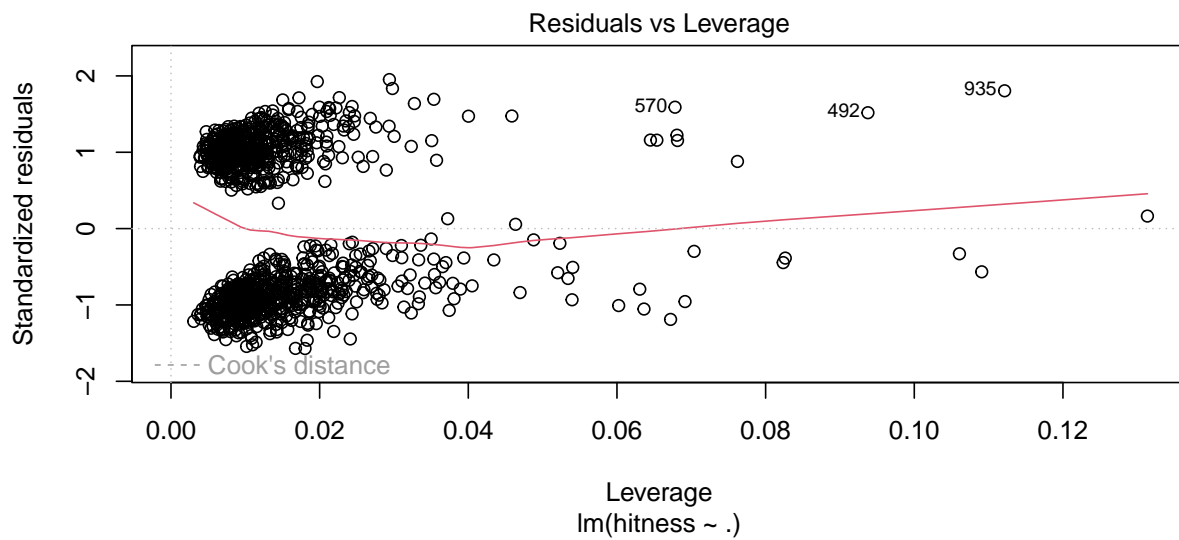
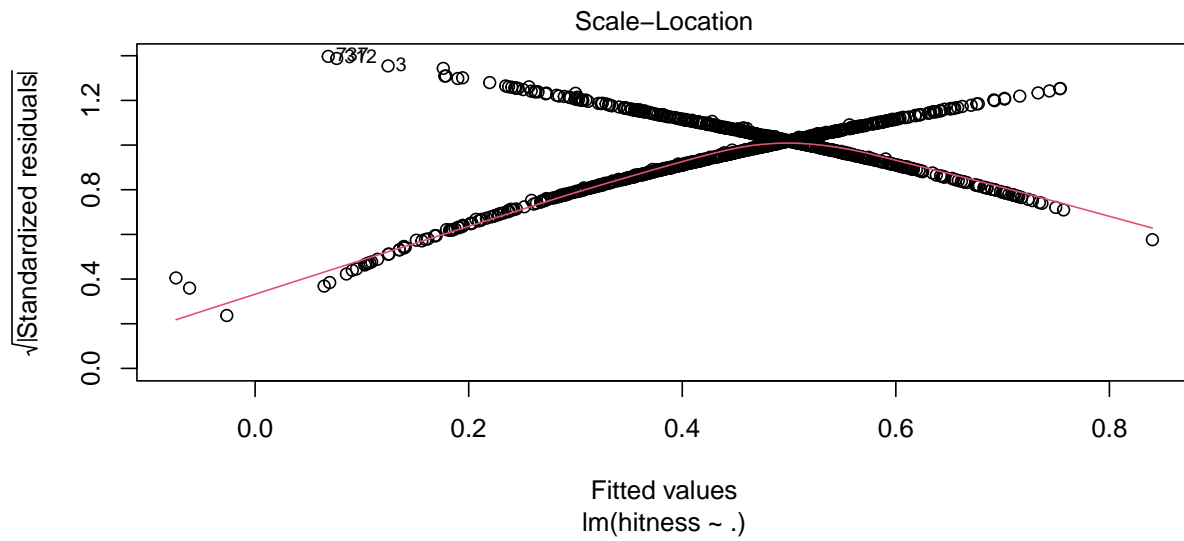
	Sum Sq	Df	F value	Pr(>F)	
danceability	1.100	1	4.6887	0.0305772	*
energy	0.445	1	1.8945	0.1689745	
key	0.031	1	0.1341	0.7142746	

loudness	0.573	1	2.4417	0.1184400	
mode	1.163	1	4.9581	0.0261717	*
speechiness	3.906	1	16.6450	4.834e-05	***
acousticness	0.825	1	3.5174	0.0609935	.
instrumentalness	0.001	1	0.0045	0.9466111	
liveness	0.002	1	0.0095	0.9224612	
valence	2.555	1	10.8864	0.0009999	***
time_signature	0.250	1	1.0660	0.3020847	
sentiment	0.421	1	1.7936	0.1807658	
duration_min	0.031	1	0.1335	0.7148837	
log_tempo	0.002	1	0.0098	0.9213401	
log_popularity	5.690	1	24.2477	9.774e-07	***
Residuals	256.949	1095			

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

Plots (Residuals vs Fitted, Normal Q-Q, Scale-Location, Residuals vs Leverage)





Logistic Regression

Accuracy

[1] 0.6147615

Summary (Coefficients, F-Statistics, P-Value)

Call:

```
glm(formula = hitness ~ ., family = binomial(logit), data = dataset.5.train)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.6732	-1.0744	-0.7175	1.1577	2.0438

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-3.55078	1.82792	-1.943	0.05207 .
danceability	1.13329	0.51531	2.199	0.02786 *
energy	-0.89314	0.64928	-1.376	0.16895
key	0.00711	0.01763	0.403	0.68673
loudness	0.06605	0.04091	1.614	0.10642
mode	0.31925	0.14052	2.272	0.02309 *
speechiness	-2.64060	0.65349	-4.041	5.33e-05 ***
acousticness	-0.69608	0.37267	-1.868	0.06179 .
instrumentalness	-0.03816	0.57944	-0.066	0.94749
liveness	0.04974	0.45602	0.109	0.91315
valence	1.12038	0.34228	3.273	0.00106 **
time_signature	0.27198	0.25577	1.063	0.28760
sentiment	0.10601	0.07879	1.346	0.17846
duration_min	0.02899	0.08147	0.356	0.72193
log_tempo	0.02621	0.26211	0.100	0.92035
log_popularity	0.21452	0.04452	4.819	1.45e-06 ***

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1531.0 on 1110 degrees of freedom
Residual deviance: 1452.5 on 1095 degrees of freedom
AIC: 1484.5

Number of Fisher Scoring iterations: 4

Anova Table

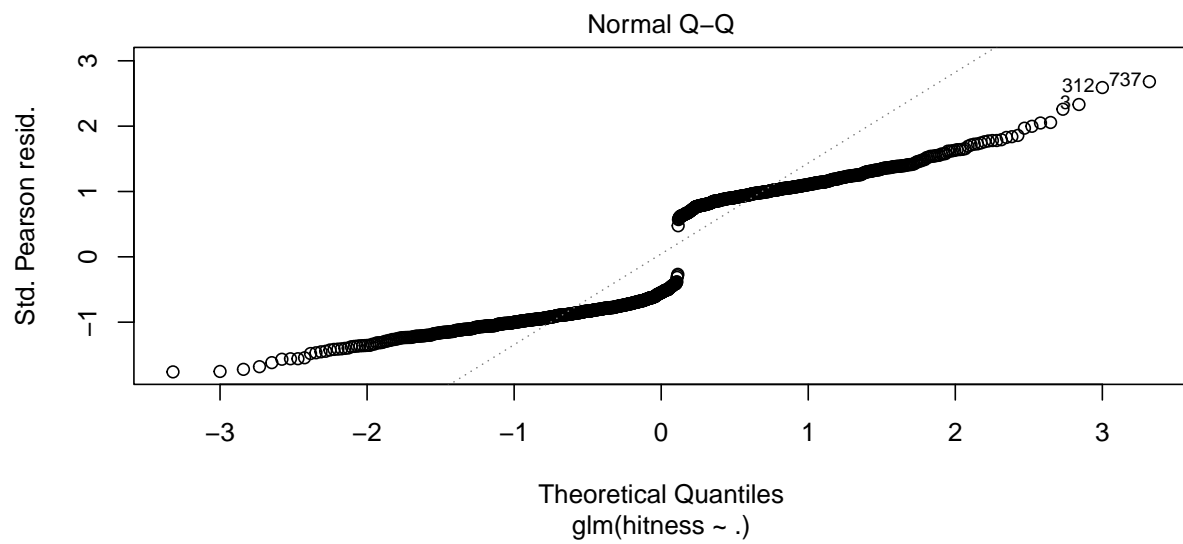
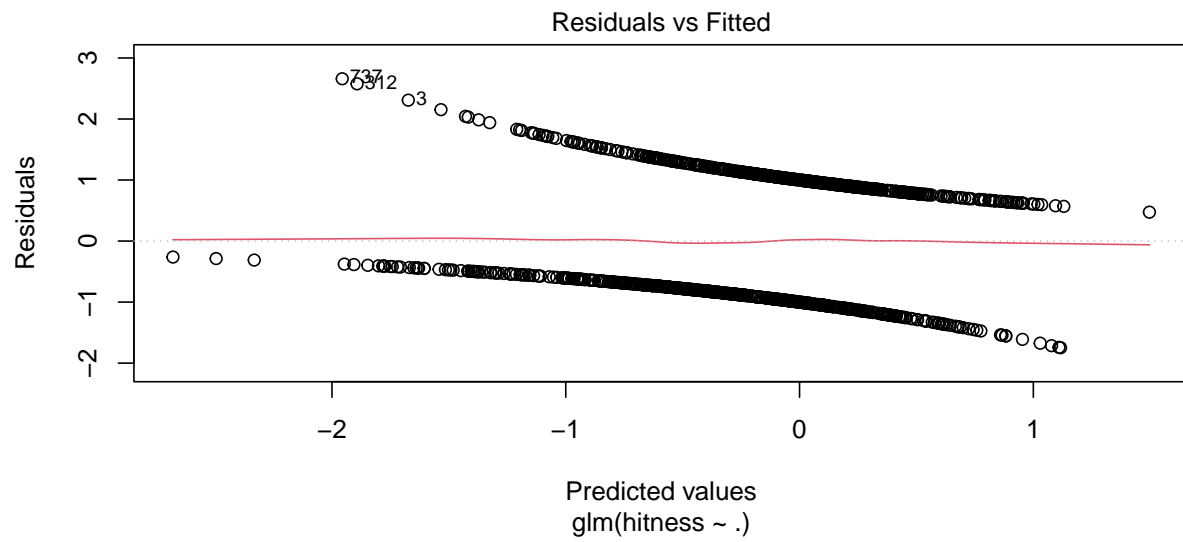
Analysis of Deviance Table (Type II tests)

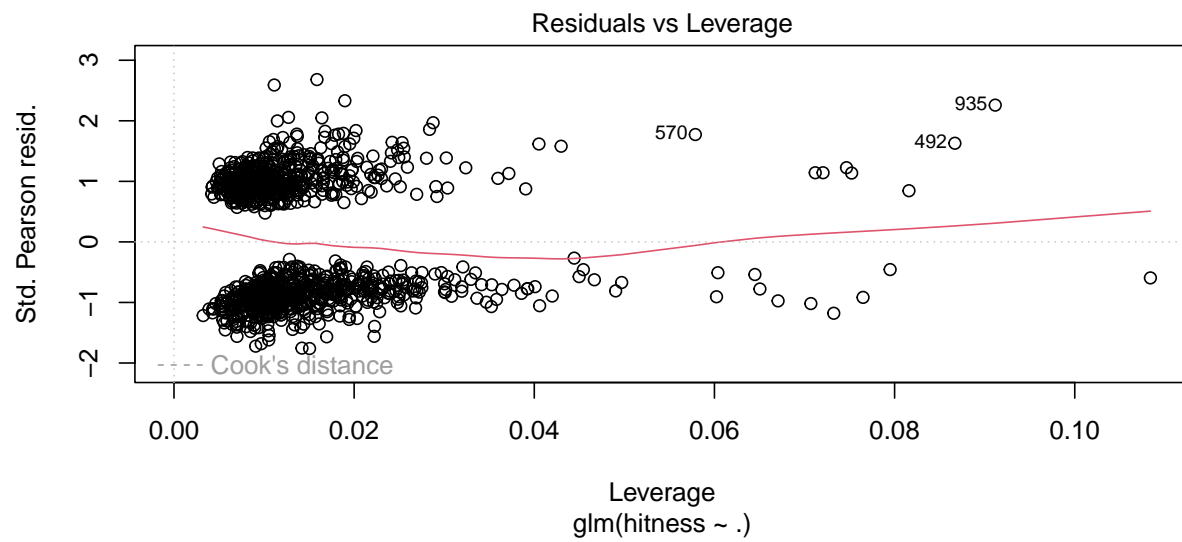
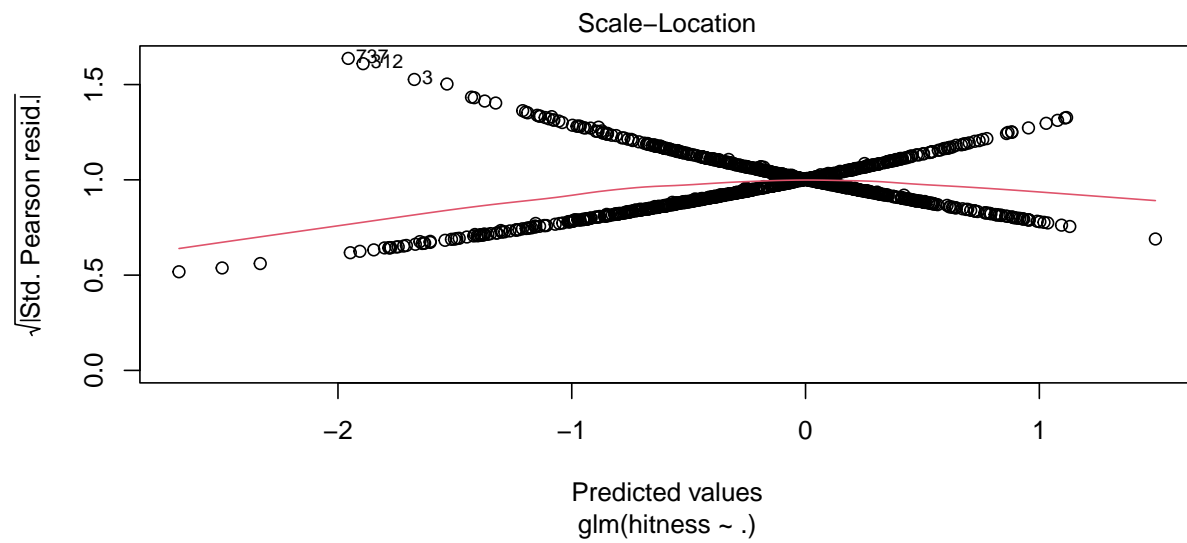
Response: hitness

	LR	Chisq	Df	Pr(>Chisq)
danceability	4.8781	1	0.027200 *	
energy	1.8988	1	0.168216	
key	0.1627	1	0.686715	
loudness	2.6252	1	0.105179	
mode	5.2005	1	0.022580 *	
speechiness	17.1550	1	3.445e-05 ***	
acousticness	3.5173	1	0.060732 .	
instrumentalness	0.0043	1	0.947450	
liveness	0.0119	1	0.913173	
valence	10.8266	1	0.001001 **	
time_signature	1.1870	1	0.275943	
sentiment	1.8156	1	0.177843	
duration_min	0.1262	1	0.722365	
log_tempo	0.0100	1	0.920341	
log_popularity	24.9659	1	5.835e-07 ***	

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

Plots (Residuals vs Fitted, Normal Q-Q, Scale-Location, Residuals vs Leverage)



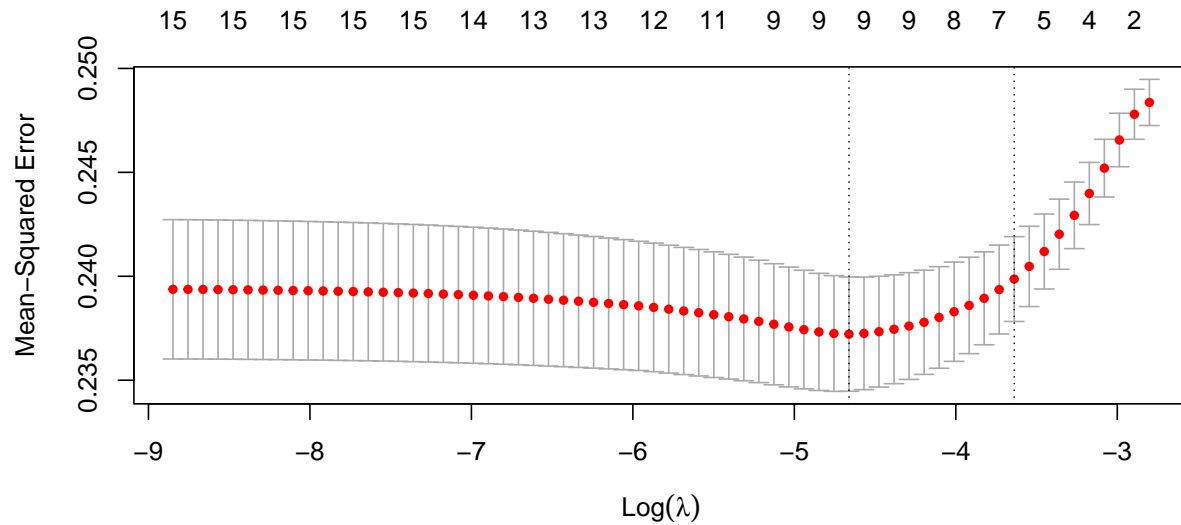


Logistic LASSO Regression

Accuracy

[1] 0.6138614

LASSO Log Selection (1SE)



Summary (Coefficients, F-Statistics, P-Value)

Call:

```
glm(formula = as.formula(paste("hitness ~", paste(fit.lasso.5.names,
collapse = " + "))), family = binomial(logit), data = dataset.5.train)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.6227	-1.0858	-0.7318	1.1605	2.0909

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-2.98807	0.52705	-5.669	1.43e-08 ***
danceability	1.23571	0.47788	2.586	0.00972 **
loudness	0.03274	0.03079	1.064	0.28753
mode	0.31169	0.13756	2.266	0.02346 *
speechiness	-2.77018	0.64117	-4.320	1.56e-05 ***
acousticness	-0.50296	0.33217	-1.514	0.12999
valence	0.97959	0.31172	3.143	0.00167 **
log_popularity	0.21701	0.04395	4.938	7.90e-07 ***

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1531.0 on 1110 degrees of freedom
Residual deviance: 1457.3 on 1103 degrees of freedom
AIC: 1473.3

Number of Fisher Scoring iterations: 4

Anova Table

Analysis of Deviance Table (Type II tests)

Response: hitness

	LR	Chisq	Df	Pr(>Chisq)
danceability	6.7522	1	0.009363	**
loudness	1.1385	1	0.285976	
mode	5.1728	1	0.022943	*
speechiness	19.7425	1	8.861e-06	***
acousticness	2.3131	1	0.128286	
valence	9.9641	1	0.001596	**
log_popularity	26.2782	1	2.956e-07	***

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

Plots (Residuals vs Fitted, Normal Q-Q, Scale-Location, Residuals vs Leverage)

