

# Analyzing Musical Feature Trends in Top Spotify Songs (2010–2019): A Statistical Overview

Jonah Sitorus

Dataset provided on Canvas (ProjectPhase2\_ExcelDataFile\_Spotify)

## Executive Summary

In this project, I analyzed data on top Spotify songs from 2010-2019 to understand what makes a song popular. The dataset includes variables on musical features such as tempo (bpm), danceability, valence (mood), duration, year, and genre, along with a popularity score from 0 to 100. Understanding these relationships helps with identifying trends in listener preferences and better support playlist and marketing decisions.

Using R, I summarize the distribution of song popularity with descriptive statistics. I explore relationships between popularity and numerical variables using scatterplots and correlations. I built a multiple regression model with numerical predictors to explain variation in popularity and test model assumptions through residual analysis. I then extended the model by adding a categorical variable for the dance pop genre to see if this genre is associated with higher and lower popularity after adjusting for other musical attributes.

Overall, the results indicate that danceability, valence, and duration are statistically significant predictors of popularity, while year and tempo are not significant in this dataset. Residual models suggested that the assumptions were reasonably met and support the validity of this model. These findings provide an insight into the attributes that are in a top-charting Spotify song and highlights which feature is most popular and what is most geared towards listener preference.

## Descriptive Statistics for Popularity

```

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
D - RStudio RStudioProjectPhase2... jst78... jst8... Unnamed... PROJECT1.R ProjectPhase2_ExcelData_Spotify... jst78... Cellphonesales...
Source
1 dataset <- spotify
2 # Load the 'readxl' library
3 library(readxl)
4 df <- ProjectPhase2_ExcelDatafile_Spotify
5
6 #5 measures of location
7 sum(df$popularity)
8 mean_spotify <- mean(df$popularity)
9 median_spotify <- median(df$popularity)
10 sd_spotify <- sd(df$popularity)
11 min_spotify <- min(df$popularity)
12 max_spotify <- max(df$popularity)
13
14 print(mean_spotify)
15 print(median_spotify)
16 print(sd_spotify)
17 print(min_spotify)
18 print(max_spotify)
19
20 #5 dispersion of popularity
21 range_spotify <- max_spotify - min_spotify
22 var_spotify <- var(df$popularity)
23 sd_spotify <- sd(df$popularity)
24 iqr_spotify <- IQR(df$popularity)
25 cv_spotify <- (sd_spotify / mean_spotify)
26
27
28 print(range_spotify)
29 print(var_spotify)
30 print(sd_spotify)
31 print(iqr_spotify)
32
33

```

Console Terminal Background Jobs

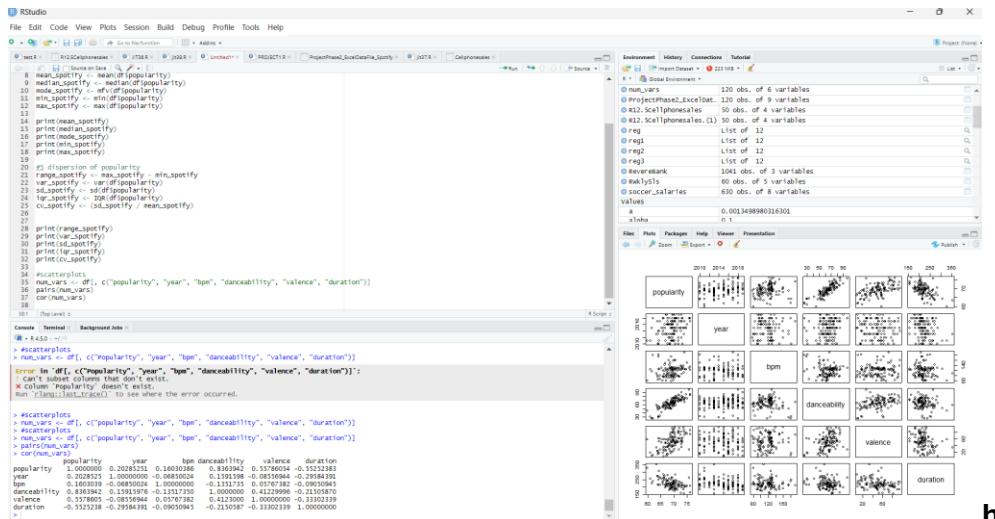
```

> R 4.0.2 -->
> var_spotify <- var(df)
warning message:
In var(df) : NAs introduced by coercion
> dispersion of popularity
> range_spotify <- max_spotify - min_spotify
> var_spotify <- var(df$popularity)
> sd_spotify <- sd(df$popularity)
> iqr_spotify <- IQR(df$popularity)
> cv_spotify <- (sd_spotify / mean_spotify)
> print(range_spotify)
[1] 16.37995
> print(var_spotify)
[1] 10.00000
> print(sd_spotify)
[1] 3.128099
> print(iqr_spotify)
[1] 3.059931
> print(cv_spotify)
[1] 0.04649345
>

```

In this section, the descriptive statistics for the dependent variable, Popularity. The mean popularity was approximately 69.19222 with a median of 69.41. The minimum and maximum values were 59.90868 and 76.28763. The standard deviation was 3.216985 which indicates that the popularity scores were not spread out that much. Because the distribution is skewed to the right this means that the mean is greater than the median and the best measure of the center is the median because it's not heavily influenced by the high-value outliers in the right tail.

## Scatterplots and Correlation



The scatterplot of Popularity versus Danceability appears to show a roughly linear pattern and the direction is positive, meaning that as danceability increases, popularity also increases. The strength of the relationships appears strong based on how the points cluster around a trend and that the correlation level is high at .836 and there's no obvious outliers.

The screenshot shows the RStudio interface. The top menu bar includes File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, and Help. The left sidebar shows several open files: test.R, R12.5Cellphonesales.R, JIT38.R, Untitled1.R, PROJECT1.R, ProjectPhase2\_ExcelDataFile\_Spotify.R, jit37.R, and Cellphonesales.R. The main workspace contains the following R code:

```

11 min_spotify <- min(df$popularity)
12 max_spotify <- max(df$popularity)
13
14 print(mean_spotify)
15 print(median_spotify)
16 print(mode_spotify)
17 print(min_spotify)
18 print(max_spotify)
19
20 #5 dispersion of popularity
21 range_spotify <- max_spotify - min_spotify
22 var_spotify <- var(df$popularity)
23 sd_spotify <- sd(df$popularity)
24 iqr_spotify <- IQR(df$popularity)
25 cv_spotify <- (sd_spotify / mean_spotify)
26
27
28 print(range_spotify)
29 print(var_spotify)
30 print(sd_spotify)
31 print(iqr_spotify)
32 print(cv_spotify)
33
34 #scatterplots
35 num_vars <- df[, c("popularity", "year", "bpm", "danceability", "valence", "duration")]
36 pairs(num_vars)
37 cor(num_vars)
38
39 #regression
40 cor.test(df$popularity, df$danceability)
41

```

The code ends with a comment `41:1 (Top Level) <-->`. Below the code, the Terminal tab shows the following output:

```

R - R 4.5.0 - ~/Desktop
Error in cor.test.default(df$Popularity, df$danceability) :
  'x' must be a numeric vector
In addition: warning message:
Unknown or uninitialized column: `Popularity`.

> #regression
> cor.test(df$popularity, df$danceability)

Pearson's product-moment correlation

data: df$popularity and df$danceability
t = 16.576, df = 118, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7730450 0.8832255
sample estimates:
cor
0.8363942
>

```

## Test significance of correlation (Popularity & Danceability)

?

**H<sub>0</sub>:**  $\rho = 0$  (no linear correlation)

?

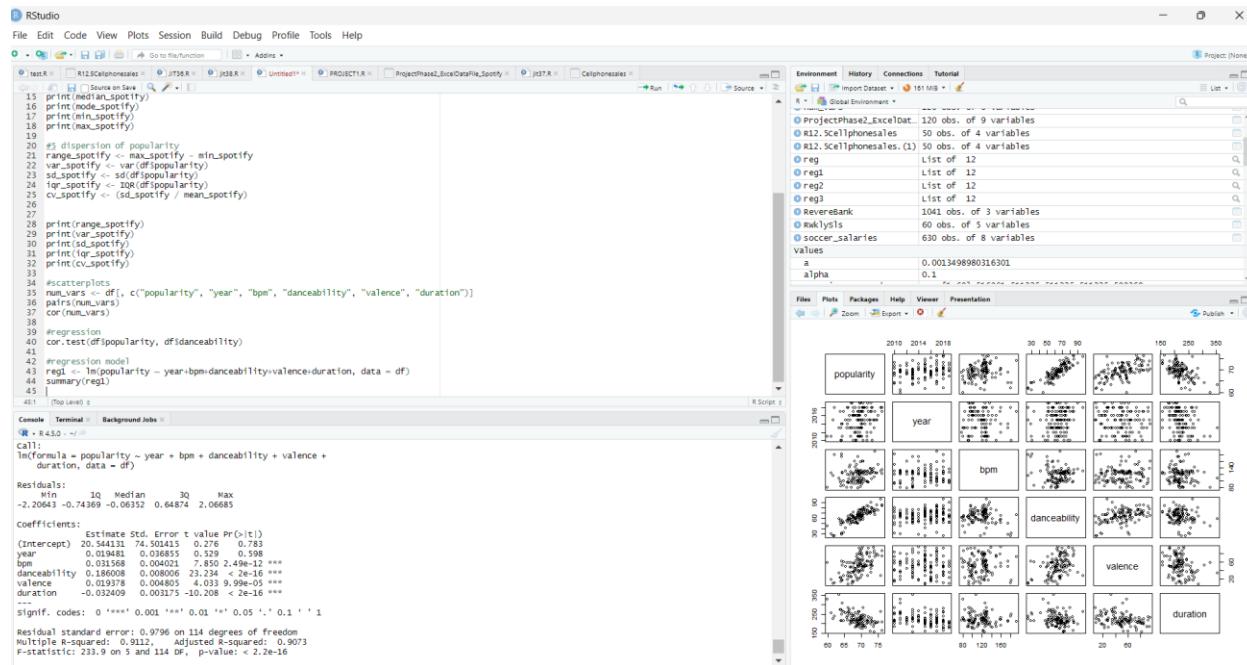
**H<sub>a</sub>:**  $\rho \neq 0$  (non-zero linear correlation)

The correlation between popularity and danceability is .836. Using `cor.test()`, we obtain a p-value of  $2.2e^{-16}$  which at the .05 significance level, we reject the null hypothesis of zero correlation. This indicates that there is a statistically significant linear association between popularity and danceability.

## Multicollinearity

For multicollinearity among the independent variables, I used the correlation matrix and pairs with correlations below 0.2 were classified as low, 0.2-0.8 as moderate and above 0.8 as extreme. In this dataset, popularity and year, popularity and bpm, popularity and duration had low multicollinearity, and popularity and valence had moderate multicollinearity and popularity and danceability and extreme multicollinearity.

## Multiple Regression Model



The r-squared for this model is 0.9112 meaning that ~91% of the variation in popularity is explained by the independent variables. This suggests the model has a strong fit to the data.

## Is R-Squared Statistically Significant?

- ?)  $H_0$ : All slope coefficients = 0 (no linear relationship between Popularity and any predictor)
- ?)  $H_a$ : At least one slope coefficient  $\neq 0$

The F-statistic is 233.9 with a p-value of 2.2e-16 and because this p-value is less than .05 we reject the null hypothesis. This means that the regression model is statistically significant in explaining the variation in popularity.

### **Conclusion and interpretation**

$\beta_0$ : if all of year, bpm, danceability, valence, and duration = 0, then estimated popularity = 20.544, not significant at the .05 level.

$\beta_1$ : year does not impact popularity with statistical significance at the .05 level.

$\beta_2$ : bpm does not impact popularity with statistical significance at the .05 level.

$\beta_3$ : for a one unit increase in danceability, pop increases by .186, holding all other variables fixed, significant at the .05 level.

$\beta_4$ : for a one unit increase in valence, pop increases by .193, holding all other variables fixed, significant at the .05 level.

$\beta_5$ : for a one unit increase in duration, pop decreases by .324, holding all other variables fixed, significant at the .05 level.

### **Residual Analysis**

To assess whether the assumptions of simple linear regression were met, several statistical tests were used. These include graphical plots as well as formal tests for normality, homoscedasticity, and independence of residuals.

### **Graphs**

RStudio

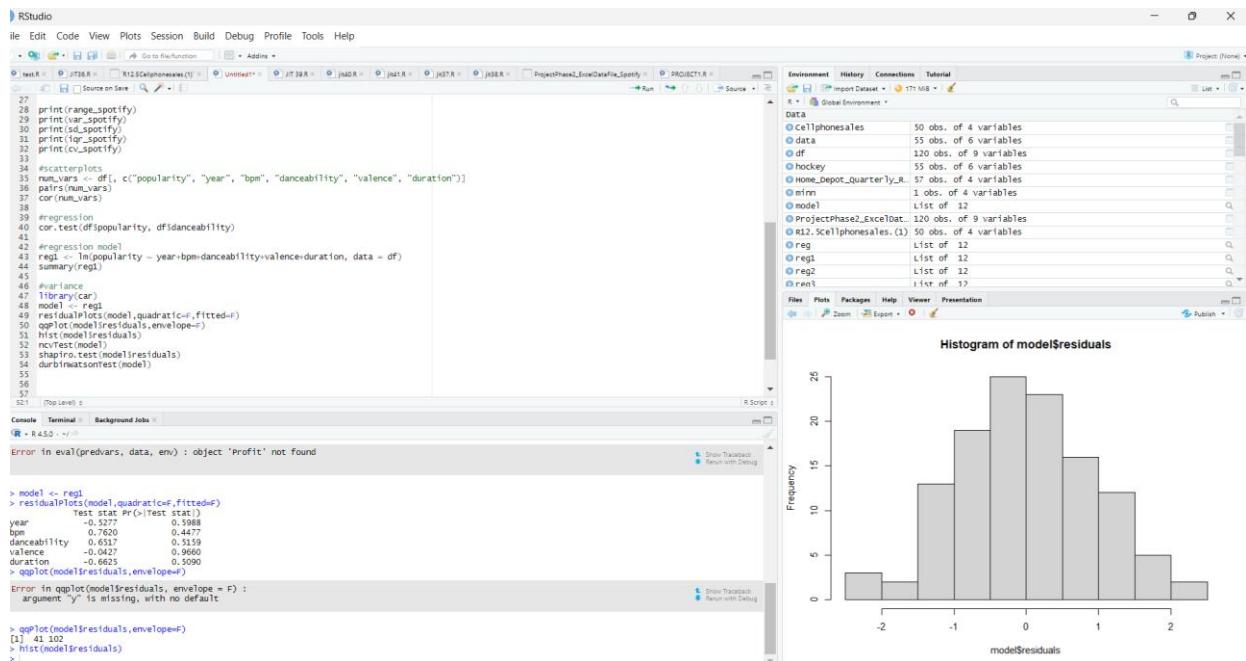
File Edit Code View Plots Session Build Debug Profile Tools Help

R Script 1

```

26 print(range_spotify)
27
28 print(var_spotify)
29 print(sd_spotify)
30 print(qr_spotify)
31 print(cv_spotify)
32
33 #scatterplots
34 num_vars <- df[, c("popularity", "year", "bpm", "danceability", "valence", "duration")]
35 pairs(num_vars)
36 cor(num_vars)
37
38 #regression
39 cor.test(popularity, df$danceability)
40
41 #regression model
42 reg1 <- lm(popularity ~ year+bpm+danceability+valence+duration, data = df)
43 summary(reg1)
44
45 #variance
46 library(car)
47 model <- reg1
48 qqplot(model$residuals, envelope=TRUE)
49 plot(model$residuals)
50 hist(model$residuals)
51 nrmn <- rnorm(57)
52 Shapiro.test(model$residuals)
53 durbinWatsonTest(model)
54
55
56
57
58
59
59.5
60
60.5
61
61.5
62
62.5
63
63.5
64
64.5
65
65.5
66
66.5
67
67.5
68
68.5
69
69.5
69.5
70
70.5
71
71.5
72
72.5
73
73.5
74
74.5
75
75.5
76
76.5
77
77.5
78
78.5
79
79.5
79.5
80
80.5
81
81.5
82
82.5
83
83.5
84
84.5
85
85.5
86
86.5
87
87.5
88
88.5
89
89.5
89.5
90
90.5
91
91.5
92
92.5
93
93.5
94
94.5
95
95.5
96
96.5
97
97.5
98
98.5
99
99.5
99.5
100
100.5
101
101.5
102
102.5
103
103.5
104
104.5
105
105.5
106
106.5
107
107.5
108
108.5
109
109.5
110
110.5
111
111.5
112
112.5
113
113.5
114
114.5
115
115.5
116
116.5
117
117.5
118
118.5
119
119.5
120
120.5
121
121.5
122
122.5
123
123.5
124
124.5
125
125.5
126
126.5
127
127.5
128
128.5
129
129.5
130
130.5
131
131.5
132
132.5
133
133.5
134
134.5
135
135.5
136
136.5
137
137.5
138
138.5
139
139.5
140
140.5
141
141.5
142
142.5
143
143.5
144
144.5
145
145.5
146
146.5
147
147.5
148
148.5
149
149.5
150
150.5
151
151.5
152
152.5
153
153.5
154
154.5
155
155.5
156
156.5
157
157.5
158
158.5
159
159.5
160
160.5
161
161.5
162
162.5
163
163.5
164
164.5
165
165.5
166
166.5
167
167.5
168
168.5
169
169.5
170
170.5
171
171.5
172
172.5
173
173.5
174
174.5
175
175.5
176
176.5
177
177.5
178
178.5
179
179.5
180
180.5
181
181.5
182
182.5
183
183.5
184
184.5
185
185.5
186
186.5
187
187.5
188
188.5
189
189.5
190
190.5
191
191.5
192
192.5
193
193.5
194
194.5
195
195.5
196
196.5
197
197.5
198
198.5
199
199.5
200
200.5
201
201.5
202
202.5
203
203.5
204
204.5
205
205.5
206
206.5
207
207.5
208
208.5
209
209.5
210
210.5
211
211.5
212
212.5
213
213.5
214
214.5
215
215.5
216
216.5
217
217.5
218
218.5
219
219.5
220
220.5
221
221.5
222
222.5
223
223.5
224
224.5
225
225.5
226
226.5
227
227.5
228
228.5
229
229.5
230
230.5
231
231.5
232
232.5
233
233.5
234
234.5
235
235.5
236
236.5
237
237.5
238
238.5
239
239.5
240
240.5
241
241.5
242
242.5
243
243.5
244
244.5
245
245.5
246
246.5
247
247.5
248
248.5
249
249.5
250
250.5
251
251.5
252
252.5
253
253.5
254
254.5
255
255.5
256
256.5
257
257.5
258
258.5
259
259.5
260
260.5
261
261.5
262
262.5
263
263.5
264
264.5
265
265.5
266
266.5
267
267.5
268
268.5
269
269.5
270
270.5
271
271.5
272
272.5
273
273.5
274
274.5
275
275.5
276
276.5
277
277.5
278
278.5
279
279.5
280
280.5
281
281.5
282
282.5
283
283.5
284
284.5
285
285.5
286
286.5
287
287.5
288
288.5
289
289.5
290
290.5
291
291.5
292
292.5
293
293.5
294
294.5
295
295.5
296
296.5
297
297.5
298
298.5
299
299.5
300
300.5
301
301.5
302
302.5
303
303.5
304
304.5
305
305.5
306
306.5
307
307.5
308
308.5
309
309.5
310
310.5
311
311.5
312
312.5
313
313.5
314
314.5
315
315.5
316
316.5
317
317.5
318
318.5
319
319.5
320
320.5
321
321.5
322
322.5
323
323.5
324
324.5
325
325.5
326
326.5
327
327.5
328
328.5
329
329.5
330
330.5
331
331.5
332
332.5
333
333.5
334
334.5
335
335.5
336
336.5
337
337.5
338
338.5
339
339.5
340
340.5
341
341.5
342
342.5
343
343.5
344
344.5
345
345.5
346
346.5
347
347.5
348
348.5
349
349.5
350
350.5
351
351.5
352
352.5
353
353.5
354
354.5
355
355.5
356
356.5
357
357.5
358
358.5
359
359.5
360
360.5
361
361.5
362
362.5
363
363.5
364
364.5
365
365.5
366
366.5
367
367.5
368
368.5
369
369.5
370
370.5
371
371.5
372
372.5
373
373.5
374
374.5
375
375.5
376
376.5
377
377.5
378
378.5
379
379.5
380
380.5
381
381.5
382
382.5
383
383.5
384
384.5
385
385.5
386
386.5
387
387.5
388
388.5
389
389.5
390
390.5
391
391.5
392
392.5
393
393.5
394
394.5
395
395.5
396
396.5
397
397.5
398
398.5
399
399.5
400
400.5
401
401.5
402
402.5
403
403.5
404
404.5
405
405.5
406
406.5
407
407.5
408
408.5
409
409.5
410
410.5
411
411.5
412
412.5
413
413.5
414
414.5
415
415.5
416
416.5
417
417.5
418
418.5
419
419.5
420
420.5
421
421.5
422
422.5
423
423.5
424
424.5
425
425.5
426
426.5
427
427.5
428
428.5
429
429.5
430
430.5
431
431.5
432
432.5
433
433.5
434
434.5
435
435.5
436
436.5
437
437.5
438
438.5
439
439.5
440
440.5
441
441.5
442
442.5
443
443.5
444
444.5
445
445.5
446
446.5
447
447.5
448
448.5
449
449.5
450
450.5
451
451.5
452
452.5
453
453.5
454
454.5
455
455.5
456
456.5
457
457.5
458
458.5
459
459.5
460
460.5
461
461.5
462
462.5
463
463.5
464
464.5
465
465.5
466
466.5
467
467.5
468
468.5
469
469.5
470
470.5
471
471.5
472
472.5
473
473.5
474
474.5
475
475.5
476
476.5
477
477.5
478
478.5
479
479.5
480
480.5
481
481.5
482
482.5
483
483.5
484
484.5
485
485.5
486
486.5
487
487.5
488
488.5
489
489.5
490
490.5
491
491.5
492
492.5
493
493.5
494
494.5
495
495.5
496
496.5
497
497.5
498
498.5
499
499.5
500
500.5
501
501.5
502
502.5
503
503.5
504
504.5
505
505.5
506
506.5
507
507.5
508
508.5
509
509.5
510
510.5
511
511.5
512
512.5
513
513.5
514
514.5
515
515.5
516
516.5
517
517.5
518
518.5
519
519.5
520
520.5
521
521.5
522
522.5
523
523.5
524
524.5
525
525.5
526
526.5
527
527.5
528
528.5
529
529.5
530
530.5
531
531.5
532
532.5
533
533.5
534
534.5
535
535.5
536
536.5
537
537.5
538
538.5
539
539.5
540
540.5
541
541.5
542
542.5
543
543.5
544
544.5
545
545.5
546
546.5
547
547.5
548
548.5
549
549.5
550
550.5
551
551.5
552
552.5
553
553.5
554
554.5
555
555.5
556
556.5
557
557.5
558
558.5
559
559.5
560
560.5
561
561.5
562
562.5
563
563.5
564
564.5
565
565.5
566
566.5
567
567.5
568
568.5
569
569.5
570
570.5
571
571.5
572
572.5
573
573.5
574
574.5
575
575.5
576
576.5
577
577.5
578
578.5
579
579.5
580
580.5
581
581.5
582
582.5
583
583.5
584
584.5
585
585.5
586
586.5
587
587.5
588
588.5
589
589.5
590
590.5
591
591.5
592
592.5
593
593.5
594
594.5
595
595.5
596
596.5
597
597.5
598
598.5
599
599.5
600
600.5
601
601.5
602
602.5
603
603.5
604
604.5
605
605.5
606
606.5
607
607.5
608
608.5
609
609.5
610
610.5
611
611.5
612
612.5
613
613.5
614
614.5
615
615.5
616
616.5
617
617.5
618
618.5
619
619.5
620
620.5
621
621.5
622
622.5
623
623.5
624
624.5
625
625.5
626
626.5
627
627.5
628
628.5
629
629.5
630
630.5
631
631.5
632
632.5
633
633.5
634
634.5
635
635.5
636
636.5
637
637.5
638
638.5
639
639.5
640
640.5
641
641.5
642
642.5
643
643.5
644
644.5
645
645.5
646
646.5
647
647.5
648
648.5
649
649.5
650
650.5
651
651.5
652
652.5
653
653.5
654
654.5
655
655.5
656
656.5
657
657.5
658
658.5
659
659.5
660
660.5
661
661.5
662
662.5
663
663.5
664
664.5
665
665.5
666
666.5
667
667.5
668
668.5
669
669.5
670
670.5
671
671.5
672
672.5
673
673.5
674
674.5
675
675.5
676
676.5
677
677.5
678
678.5
679
679.5
680
680.5
681
681.5
682
682.5
683
683.5
684
684.5
685
685.5
686
686.5
687
687.5
688
688.5
689
689.5
690
690.5
691
691.5
692
692.5
693
693.5
694
694.5
695
695.5
696
696.5
697
697.5
698
698.5
699
699.5
700
700.5
701
701.5
702
702.5
703
703.5
704
704.5
705
705.5
706
706.5
707
707.5
708
708.5
709
709.5
710
710.5
711
711.5
712
712.5
713
713.5
714
714.5
715
715.5
716
716.5
717
717.5
718
718.5
719
719.5
720
720.5
721
721.5
722
722.5
723
723.5
724
724.5
725
725.5
726
726.5
727
727.5
728
728.5
729
729.5
730
730.5
731
731.5
732
732.5
733
733.5
734
734.5
735
735.5
736
736.5
737
737.5
738
738.5
739
739.5
740
740.5
741
741.5
742
742.5
743
743.5
744
744.5
745
745.5
746
746.5
747
747.5
748
748.5
749
749.5
750
750.5
751
751.5
752
752.5
753
753.5
754
754.5
755
755.5
756
756.5
757
757.5
758
758.5
759
759.5
760
760.5
761
761.5
762
762.5
763
763.5
764
764.5
765
765.5
766
766.5
767
767.5
768
768.5
769
769.5
770
770.5
771
771.5
772
772.5
773
773.5
774
774.5
775
775.5
776
776.5
777
777.5
778
778.5
779
779.5
780
780.5
781
781.5
782
782.5
783
783.5
784
784.5
785
785.5
786
786.5
787
787.5
788
788.5
789
789.5
790
790.5
791
791.5
792
792.5
793
793.5
794
794.5
795
795.5
796
796.5
797
797.5
798
798.5
799
799.5
800
800.5
801
801.5
802
802.5
803
803.5
804
804.5
805
805.5
806
806.5
807
807.5
808
808.5
809
809.5
810
810.5
811
811.5
812
812.5
813
813.5
814
814.5
815
815.5
816
816.5
817
817.5
818
818.5
819
819.5
820
820.5
821
821.5
822
822.5
823
823.5
824
824.5
825
825.5
826
826.5
827
827.5
828
828.5
829
829.5
830
830.5
831
831.5
832
832.5
833
833.5
834
834.5
835
835.5
836
836.5
837
837.5
838
838.5
839
839.5
840
840.5
841
841.5
842
842.5
843
843.5
844
844.5
845
845.5
846
846.5
847
847.5
848
848.5
849
849.5
850
850.5
851
851.5
852
852.5
853
853.5
854
854.5
855
855.5
856
856.5
857
857.5
858
858.5
859
859.5
860
860.5
861
861.5
862
862.5
863
863.5
864
864.5
865
865.5
866
866.5
867
867.5
868
868.5
869
869.5
870
870.5
871
871.5
872
872.5
873
873.5
874
874.5
875
875.5
876
876.5
877
877.5
878
878.5
879
879.5
880
880.5
881
881.5
882
882.5
883
883.5
884
884.5
885
885.5
886
886.5
887
887.5
888
888.5
889
889.5
890
890.5
891
891.5
892
892.5
893
893.5
894
894.5
895
895.5
896
896.5
897
897.5
898
898.5
899
899.5
900
900.5
901
901.5
902
902.5
903
903.5
904
904.5
905
905.5
906
906.5
907
907.5
908
908.5
909
909.5
910
910.5
911
911.5
912
912.5
913
913.5
914
914.5
915
915.5
916
916.5
917
917.5
918
918.5
919
919.5
920
920.5
921
921.5
922
922.5
923
923.5
924
924.5
925
925.5
926
926.5
927
927.5
928
928.5
929
929.5
930
930.5
931
931.5
932
932.5
933
933.5
934
934.5
935
935.5
936
936.5
937
937.5
938
938.5
939
939.5
940
940.5
941
941.5
942
942.5
943
943.5
944
944.5
945
945.5
946
946.5
947
947.5
948
948.5
949
949.5
950
950.5
951
951.5
952
952.5
953
953.5
954
954.5
955
955.5
956
956.5
957
957.5
958
958.5
959
959.5
960
960.5
961
961.5
962
962.5
963
963.5
964
964.5
965
965.5
966
966.5
967
967.5
968
968.5
969
969.5
970
970.5
971
971.5
972
972.5
973
973.5
974
974.5
975
975.5
976
976.5
977
977.5
978
978.5
979
979.5
980
980.5
981
981.5
982
982.5
983
983.5
984
984.5
985
985.5
986
986.5
987
987.5
988
988.5
989
989.5
990
990.5
991
991.5
992
992.5
993
993.5
994
994.5
995
995.5
996
996.5
997
997.5
998
998.5
999
999.5
999.5
1000
1000.5
1001
1001.5
1002
1002.5
1003
1003.5
1004
1004.5
1005
1005.5
1006
1006.5
1007
1007.5
1008
1008.5
1009
1009.5
1010
1010.5
1011
1011.5
1012
1012.5
1013
1013.5
1014
1014.5
1015
1015.5
1016
1016.5
1017
1017.5
1018
1018.5
1019
1019.5
1020
1020.5
1021
1021.5
1022
1022.5
1023
1023.5
1024
1024.5
1025
1025.5
1026
1026.5
1027
1027.5
1028
1028.5
1029
1029.5
1030
1030.5
1031
1031.5
1032
1032.5
1033
1033.5
1034
1034.5
1035
1035.5
1036
1036.5
1037
1037.5
1038
1038.5
1039
1039.5
1040
1040.5
1041
1041.5
1042
1042.5
1043
1043.5
1044
1044.5
1045
1045.5
1046
1046.5
1047
1047.5
1048
1048.5
1049
1049.5
1050
1050.5
1051
1051.5
1052
1052.5
1053
1053.5
1054
1054.5
1055
1055.5
1056
1056.5
1057
1057.5
1058
1058.5
1059
1059.5
1060
1060.5
1061
1061.5
1062
1062.5
1063
1063.5
1064
1064.5
1065
1065.5
1066
1066.5
1067
1067.5
1068
1068.5
1069
1069.5
1070
1070.5
1071
1071.5
1072
1072.5
1073
1073.5
1074
1074.5
1075
1075.5
1076
1076.5
1077
1077.5
1078
1078.5
1079
1079.5
1080
1080.5
1081
1081.5
1082
1082.5
1083
1083.5
1084
1084.5
1085
1085.5
1086
1086.5
1087
1087.5
1088
1088.5
1089
1089.5
1090
1090.5
1091
1091.5
1092
1092.5
1093
1093.5
1094
1094.5
1095
1095.5
1096
1096.5
1097
1097.5
1098
1098.5
1099
1099.5
1099.5
1100
1100.5
1101
1101.5
1102
1102.5
1103
1103.5
1104
1104.5
1105
1105.5
1106
1106.5
1107
1107.5
1108
1108.5
1109
1109.5
1110
1110.5
1111
1111.5
1112
1112.5
1113
1113.5
1114
1114.5
1115
1115.5
1116
1116.5
1117
1117.5
1118
1118.5
1119
1119.5
1120
1120.5
1121
1121.5
1122
1122.5
1123
1123.5
1124
1124.5
1125
1125.5
1126
1126.5
1127
1127.5
1128
1128.5
1129
1129.5
1130
1130.5
1131
1131.5
1132
1132.5
1133
1133.5
1134
1134.5
1135
1135.5
1136
1136.5
1137
1137.5
1138
1138.5
1139
1139.5
1140
1140.5
1141
1141.5
1142
1142.5

```



The residual plot reveals a clear pattern in year and its residual value and it slowly gets random in the other independent variables. This could mean that the linear assumption is met and that there's evidence for homoscedasticity if the plot shows a clear pattern. For the Q-Q plot, there is a strong linear pattern in the graph which could indicate normal linearity. The histogram is unimodal and approximately symmetric, indicating a normal distribution of the data.

## Breusch-Pagan Test

The screenshot shows the RStudio interface. The top part is the R Script editor with the following code:

```

27 print(range_spotify)
28 print(var_spotify)
29 print(sd_spotify)
30 print(iqr_spotify)
31 print(cv_spotify)
32
33
34 #scatterplots
35 num_vars <- df[, c("popularity", "year", "bpm", "danceability", "valence", "duration")]
36 pairs(num_vars)
37 cor(num_vars)
38
39 #regression
40 cor.test(df$popularity, df$danceability)
41
42 #regression model
43 reg1 <- lm(popularity ~ year+bpm+danceability+valence+duration, data = df)
44 summary(reg1)
45
46 #variance
47 library(car)
48 model <- reg1
49 residualPlots(model,quadratic=F,fitted=F)
50 qqplot(model$residuals, envelope=F)
51 hist(model$residuals)
52 ncvTest(model)
53 shapiro.test(model$residuals)
54 durbinWatsonTest(model)
55
56
57

```

The bottom part is the R Console window showing the results of the regression analysis:

```

> valence      -0.0427      0.9660
> duration     -0.6625      0.5090
> qqplot(model$residuals, envelope=F)
Error in qqplot(model$residuals, envelope = F) :
  argument "y" is missing, with no default

> qqplot(model$residuals, envelope=F)
[1] 41 102
> hist(model$residuals)
> residualPlots(model,quadratic=F,fitted=F)
  Test stat Pr(>|Test stat|)
year      -0.5277    0.5988
bpm       0.7620    0.4477
danceability  0.6517    0.5159
valence   -0.0427    0.9660
duration   -0.6625    0.5090
> ncvTest(model)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
chisquare = 2.615788, Df = 1, p = 0.1058
>

```

Interpretation:

- If  $p > 0.05 \rightarrow$  Homoscedasticity holds (constant variance).
- If  $p < 0.05 \rightarrow$  Heteroscedasticity present.
- The Breusch–Pagan test resulted in  $p > 0.05$ , suggesting the variance of residuals is constant, and we fail to reject that homoscedasticity is present.

## Shapiro-Wilk Test

The screenshot shows the RStudio interface. The top menu bar includes File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Help. Below the menu is a toolbar with various icons. The main area has tabs for 'test.R', 'JIT36.R', 'R12.5Cellphonesales(1)', 'Untitled1\*', 'JIT 39.R', 'JIT40.R', 'JIT41.R', 'JIT37.R', 'JIT38.R', 'ProjectPhase2\_ExcelDataFile\_Spotify', and 'PROJECT1.R'. The code editor window contains the following R script:

```

32 print(cv_spotify)
33
34 #scatterplots
35 num_vars <- df[, c("popularity", "year", "bpm", "danceability", "valence", "duration")]
36 pairs(num_vars)
37 cor(num_vars)
38
39 #regression
40 cor.test(df$popularity, df$danceability)
41
42 #regression model
43 reg1 <- lm(popularity ~ year+bpm+danceability+valence+duration, data = df)
44 summary(reg1)
45
46 #variance
47 library(car)
48 model <- reg1
49 residualPlots(model,quadratic=F,fitted=F)
50 qqplot(model$residuals,envelope=F)
51 hist(model$residuals)
52 ncvTest(model)
53 shapiro.test(model$residuals)
54 durbinWatsonTest(model)
55
56
57
58 #new variable
59 df$dancepop <- ifelse(df$genre == "dance pop", 1, 0)
60 reg2 <- lm(popularity ~ year + bpm + danceability + valence + duration + dancepop, data = df)
61 summary(reg2)
62

```

The 'Console' tab shows the following output:

```

> qqplot(model$residuals,envelope=F)
[1] 41 102
> hist(model$residuals)
> residualPlots(model,quadratic=F,fitted=F)
  Test stat Pr(>|Test stat|)
year      -0.5277    0.5988
bpm       0.7620    0.4477
danceability  0.6517    0.5159
valence   -0.0427    0.9660
duration  -0.6625    0.5090
> ncvTest(model)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chi-square = 2.615788, Df = 1, p = 0.1058
> shapiro.test(model$residuals)

Shapiro-Wilk normality test

data: model$residuals
W = 0.99059, p-value = 0.588
> |

```

### Interpretation:

- If  $p > 0.05 \rightarrow$  Fail to reject normality
- If  $p < 0.05 \rightarrow$  Evidence of non-normality.
- The Shapiro–Wilk test yielded  $p > 0.05$ , indicating that we fail to reject normality
- The result aligns with what is seen visually in the Q–Q plot and histogram.

### Durbin–Watson Test

The screenshot shows the RStudio interface. The top menu bar includes File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, and Help. The main area displays an R script with several lines of code. The code includes data loading, correlation analysis, regression modeling, and diagnostic tests like Durbin-Watson and Shapiro-Wilk. Below the script is a console window showing the output of the executed commands.

```

32 print(cv_spotify)
33
34 #scatterplots
35 num_vars <- df[, c("popularity", "year", "bpm", "danceability", "valence", "duration")]
36 pairs(num_vars)
37 cor(num_vars)
38
39 #regression
40 cor.test(df$popularity, df$danceability)
41
42 #regression model
43 reg1 <- lm(popularity ~ year+bpm+danceability+valence+duration, data = df)
44 summary(reg1)
45
46 #variance
47 library(car)
48 model <- reg1
49 residualPlots(model,quadratic=F,fitted=F)
50 qqplot(model$residuals, envelope=F)
51 hist(model$residuals)
52 ncvTest(model)
53 shapiro.test(model$residuals)
54 durbinWatsonTest(model)
55
56
57
58 #new variable
59 df$dancepop <- ifelse(df$genre == "dance pop", 1, 0)
60 reg2 <- lm(popularity ~ year + bpm + danceability + valence + duration + dancepop, data = df)
61 summary(reg2)
62

```

Console output:

```

> qqPlot(model$residuals, envelope=F)
[1] 41 102
> hist(model$residuals)
> residualPlots(model,quadratic=F,fitted=F)
  Test stat Pr(>|Test stat|)
year      -0.5277    0.5988
bpm       0.7620    0.4477
danceability  0.6517    0.5159
valence     -0.0427    0.9660
duration    -0.6625    0.5090
> ncvTest(model)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 2.615788, df = 1, p = 0.1058
> shapiro.test(model$residuals)

Shapiro-Wilk normality test

data: model$residuals
W = 0.99059, p-value = 0.588
> |

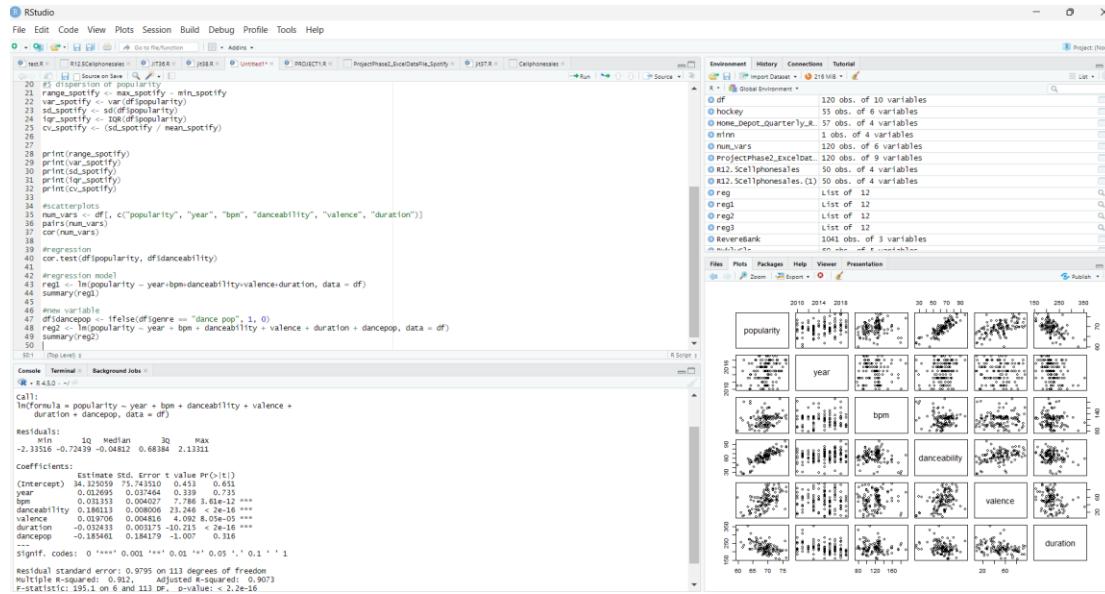
```

4

## Interpretation:

- $p > 0.05 \rightarrow$  No significant autocorrelation (independent errors)
- $p < 0.05 \rightarrow$  Evidence of autocorrelation
- The Durbin–Watson test produced  $p > 0.05$ , indicating that there is no significant autocorrelation
- This suggests the independence assumption is met.

All diagnostic tools and statistical tests indicate that the regression assumptions are satisfied and so the regression model provides a reliable and unbiased fit for the data.



## Extended Model Including Genre

The adjusted r-squared for the original model was .9073 and when the dancepop variable was added the adjusted r-squared became .9073. Because the adjusted r-squared was the same it didn't improve the model's explanatory power.

$\beta_6$ : The coefficient on dance pop is -.1854 with a p-value of .316. This means that with all the variables held constant, songs classified as dance pop are predicted to have .1854 less popularity points than non-dance pop songs. The p-value is greater than 0.05, this is statistically significant at the .05 level.