**Blackbird Records**

**Portfolio Option 1**

Krista O’Neill

CSU Global

MIS500: Foundations of Data Analytics

Dr. Kimberly Ford

May 8, 2020

Blackbird Records

Music is an integral part of human culture and society. The desire to create and to share music has been fostered throughout humanity for many ages, as an invaluable expression of art. Humans have explored pitch, tone, tempo, and harmony for years, amassing a huge amount of genres and subgenres from which to select.

Though general musical creativity should never be limited by science, in the modern-day music scene, companies need to make sure that their investment turns a profit, and sometimes that means tailoring their product for the public. Blackbird Records, a small fictional record label, is a startup company which wishes to make sure that when they mix songs, they know what factors are the most important in making a song popular.

They have asked for an analysis of Amazon’s Million Song dataset, which has given each song a *hotttnesss* score between 0-1 according to its popularity upon download, in 2010 (Million Song Dataset, n.d.). They wish to answer the questions: *1) Do tempo or duration of a song significantly impact a song’s popularity? 2) Which structural variables of a song affect its popularity the most?*

**Dataset Information**

The Million Song Dataset is published by The Echo Nest in collaboration with LabRosa (Million Song Dataset, 2020). Data was collected by sampling audio tracks, and analyzing the samples in a variety of categories, including tempo, loudness, song key, time signature, and more.

A subset is available for download as well, containing information from 10,000 songs; this subset was used for the purposes of the following analysis. The file analyzed for Blackbird Records was obtained via Github user Vatshayan (2018), who unpacked the tar.gz file for use as a .csv.

**Data Management, Storage, and Security**

Because this musical data is public access and contains objective, non-confidential information, and is also managed by The Echo Nest, Blackbird Records does not need to worry about data storage, security, or management. The Echo Nest has made the dataset available to all individuals and enterprises who wish to use the information for research and analysis. It can be stored anywhere, and queried with any free-access data analysis program.

This is beneficial to Blackbird Records since they are a starting company; there are no associated fees when it comes to servers or hiring database systems managers. If they wanted to, they could also collect data from songs they produce, and compare the songs they have produced with the pre-existing records in the database.

**Analysis Plan**

Blackbird Records wants to see if tempo or duration have an effect on song hotttnesss. Therefore, by doing two-sample t-tests using samples containing contrasting examples of the variable, it can be determined if those variables are significant. Finally, a multiple linear regression can show which song-structure variables impact song popularity the most.

**Expected Results**

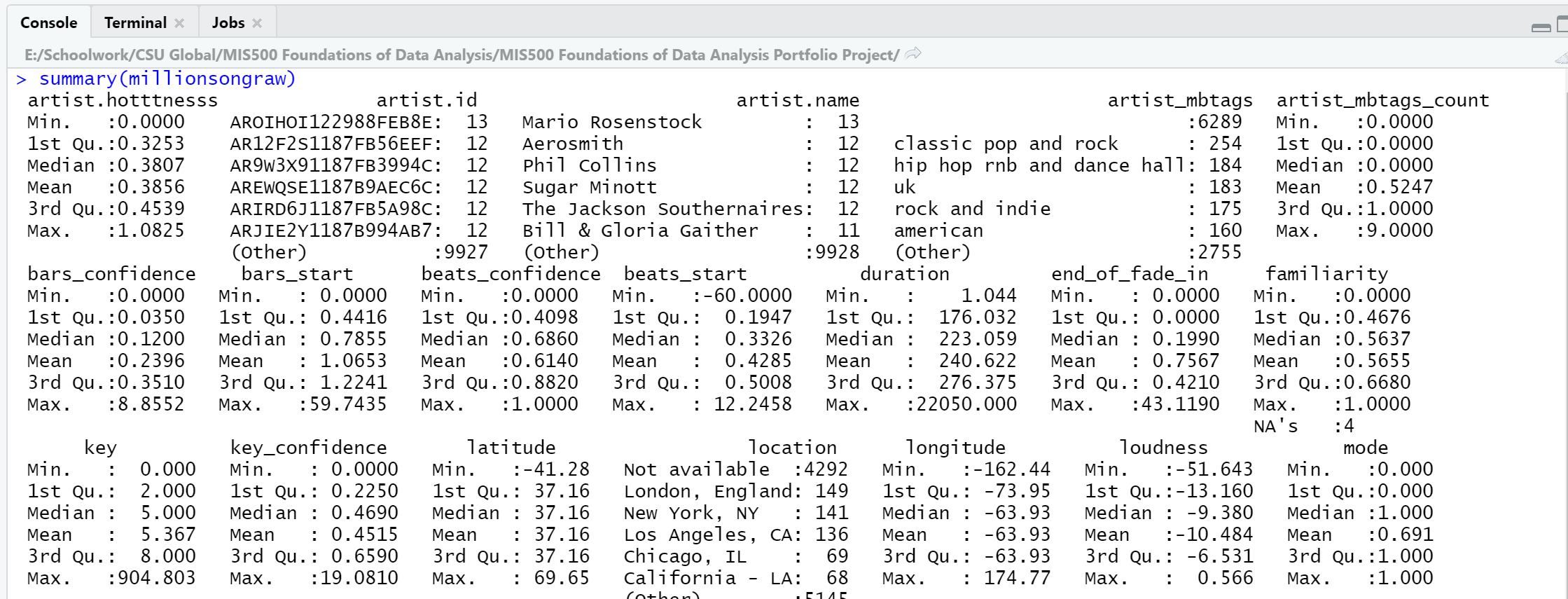
Because music is closely linked with human memory and emotions, it is natural to expect that the popularity of songs may depend strongly on variables like tempo. An upbeat tempo might be indicative of a more popular song, as upbeat tempos often express emotions like happiness or even anger, while slower tempos might indicate sorrow. Research has also shown that a preference in tempo might be usable in identifying personality traits (Dobrota & Ercegovac, 2015). Tempo has also been linked to distinct neurological responses that can affect emotions (Liu et. al, 2018).

It seems unlikely that song duration really affects a song’s popularity, unless the song were extremely short or extremely long. The #1 song for the week of May 2, 2020, according to Billboard’s Hot 100, clocks in at 3 minutes, 23 seconds, while the #2 song for the same week is 5 minutes, 3 seconds long (Billboard, 2019).

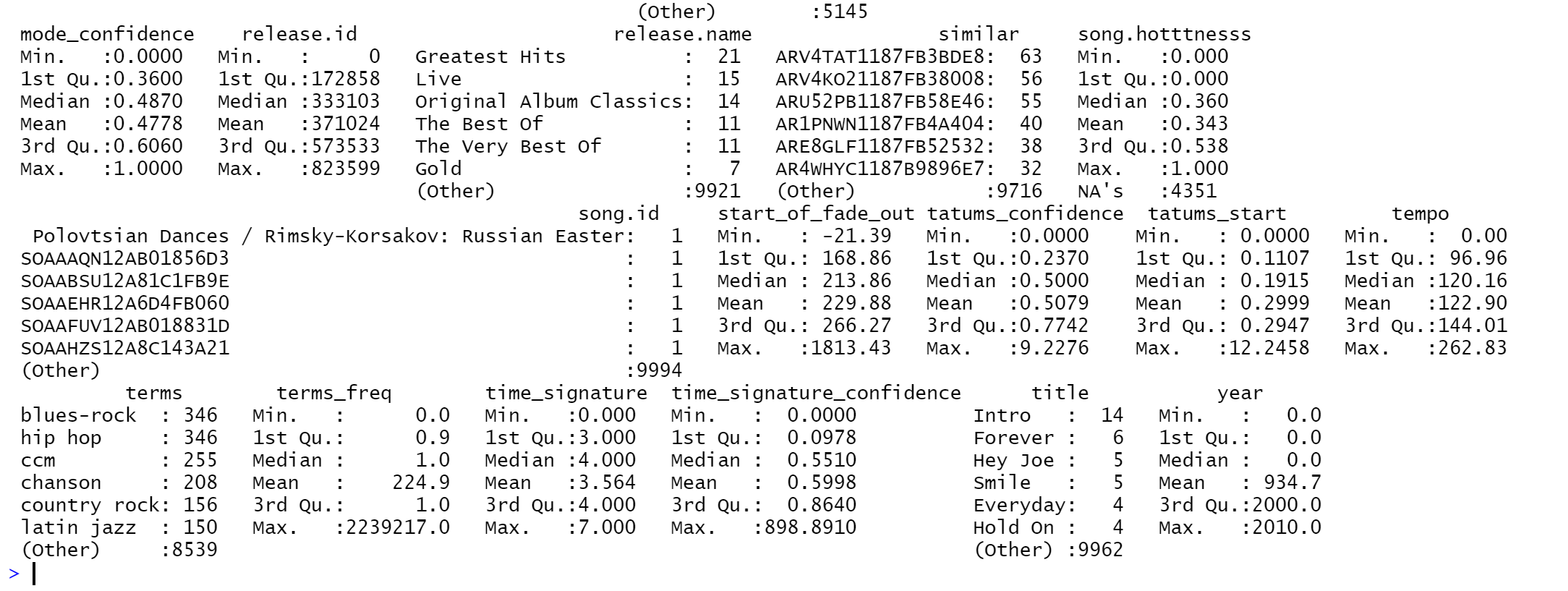
When it comes to answering the second question Blackbird Records has posed, it is highly likely that a variable like loudness would affect song popularity. Additionally, time signature is mathematical and has to do with song structure and rhythm, which are key components of music. It is likely that time signature affects song hotttnesss as well.

The only variable which might not be too significant is song key, which just affects the pitches at which the music is played. However, a closer analysis between major and minor keys might reveal that major keys might be favored in popularity over minor keys, or vice versa.

**Data Summary & Cleaning**

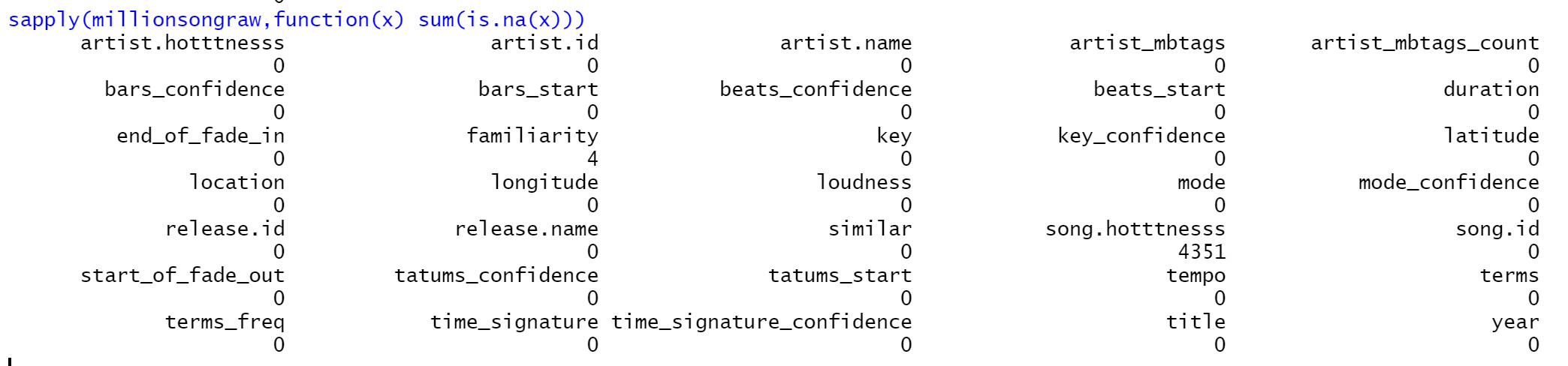


*Figure 1.* A snapshot of the dataset summary.

****

*Figure 2*. A second snapshot of the dataset summary, this time including the variable *song*.*hotttness*

A summary of the dataset revealed that there were 4,351 songs that did not have a hotttnesss value, so those values were removed from the total dataset. There are four observations missing the variable $familiarity, which isn’t important since the company’s interest is in music artists that are not previously known.



*Figure 3*. Showing missing values in the dataset.

There were 10 songs with tempo values of 0.00, which is impossible, so they were removed from the dataset as well.

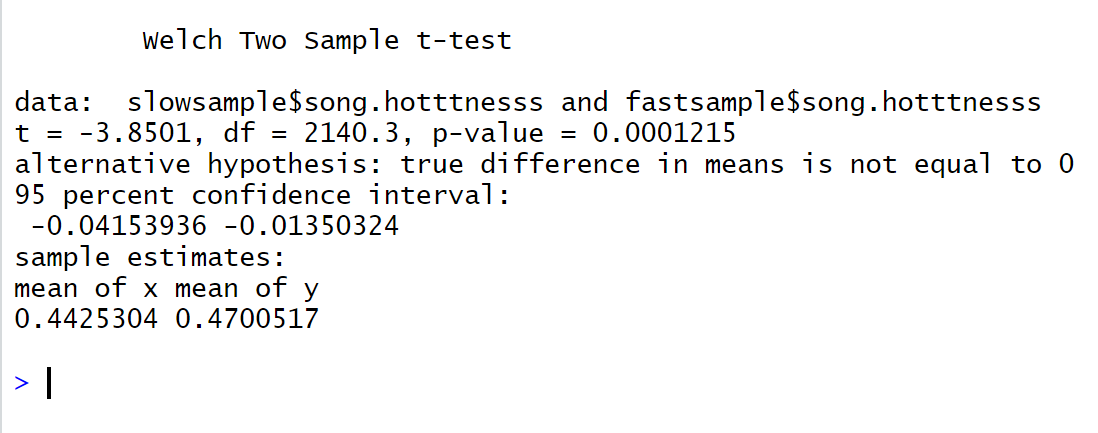
If a song’s hotttness is listed as 0, that means there was no calculation for its rating whatsoever, so those observations were also removed.

**Tempo T-Test**

A summary of the tempo variable reveals that the minimum tempo is 1 and the first quartile is 97.97 bpm, while the third quartile tempo is 146.46 and the maximum is 258.68.

To prepare for an unpaired two-sample t-test, two samples were created, Sample 1 with songs whose tempos are between the minimum and Q1, and Sample 2 with songs whose tempos are between Q3 and the maximum. Sample 1 contains 1,011 observations; Sample 2 contains 1,090. All variables except for $song.hotttnesss and $tempo were removed for this test.

The H0 of the test is, “The mean of the low-tempo sample is not significantly different than the mean of the high-tempo sample.” The H1 of the test is, “The means of the low-tempo and high-tempo sample are significantly different.”



*Figure 4.* Showing the unpaired two-sample t-test for Sample 1, with slow tempos, vs. Sample 2, with fast tempos.

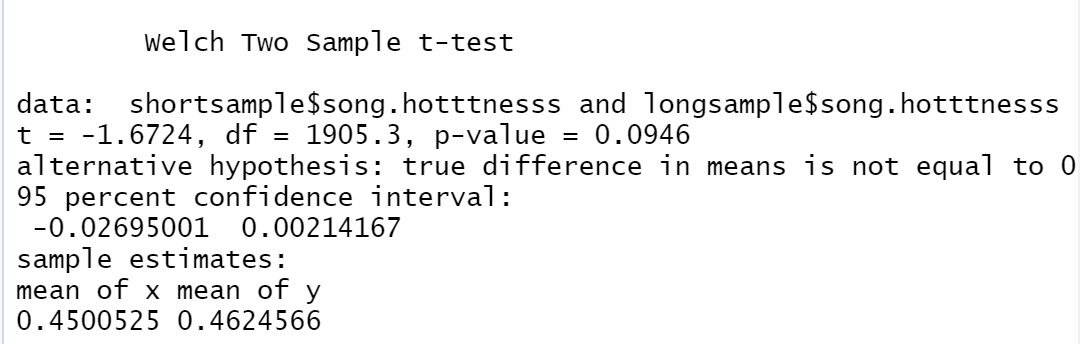
The t-test’s significance value is much smaller than 0.05, showing that H0 must be rejected and H1 is accepted. It appears that the difference in average hotttnesss of the high-tempo sample and low-tempo sample is roughly 0.47 - 0.443 = 0.027. Because the song hotttnesss scale is between 0 and 1, this is equivalent to a 2.7% difference.

**Duration T-Test**

A summary of the millionsong dataset shows that there is a maximum value of 1686.752 seconds. Because the mean is 240.69, this is an outlier. The standard deviation of the variable $duration is 103.1178, so all values that fall 2 standard deviations above the mean (so above 446.9256) are removed before proceeding with the t-test.

Two samples were created. Sample 1 contains songs with durations within the 1st quartile, so durations of less than 184.85 seconds. Sample 2 contains songs with durations above the 3rd quartile and below 2 standard deviations above the mean, so less than 447.0856 seconds.

H0: The mean $song.hotttnesss of Sample 1 is not significantly different than the mean $song.hotttnesss of Sample 2. H1: The means of the samples’ $song.hotttnesss are significantly different.



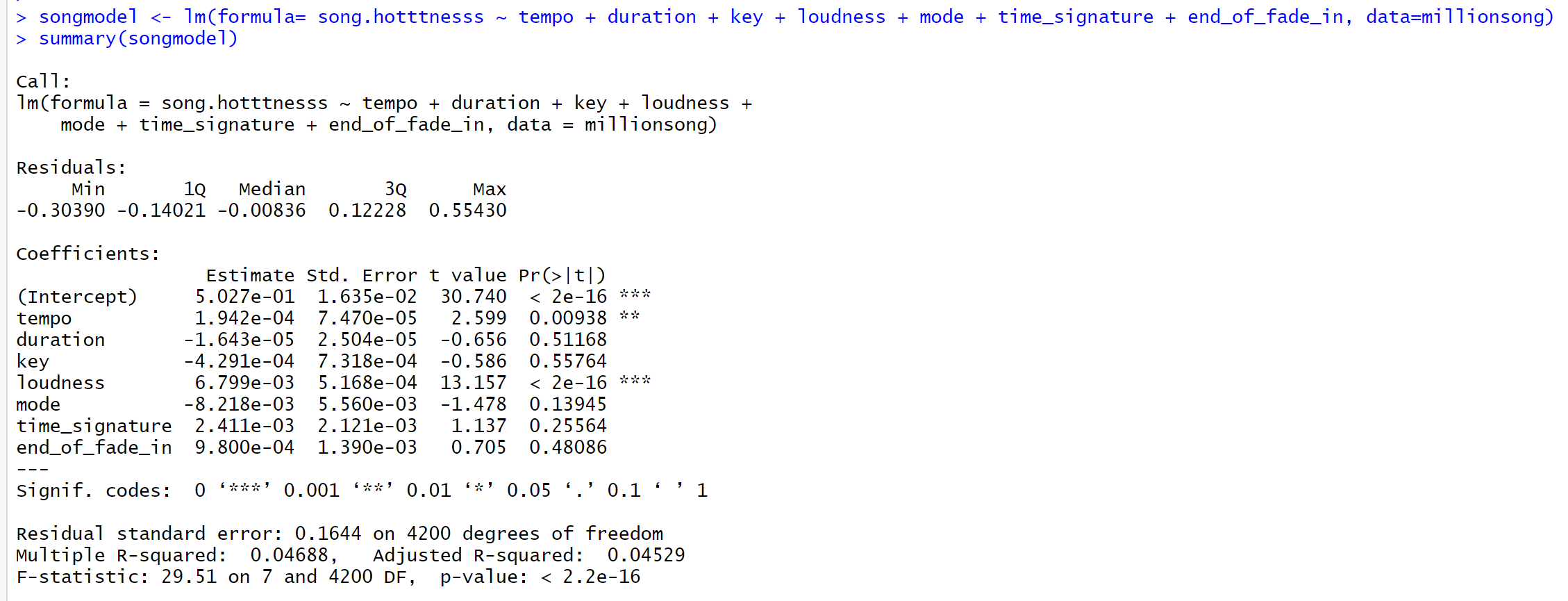
*Figure 5.* The duration t-test. The p-value is greater than 0.05, showing that there is no significant difference between the means.

Because the p-value is above 0.05, the alternative hypothesis must be rejected, and it can be stated that there is no significant difference between the means of songs with short durations or longer durations. Therefore, duration is not an important factor when it comes to mixing a song with the goal of making it as popular as possible.

**Finding Significant Variables**

To determine which variables are most significant, it is helpful to employ a linear model. Because the value $song.hotttnesss is not a binary value, a simple linear model is sufficient. Variables which have to do with the song’s overall structural characteristics were used in the model: tempo, duration, key, loudness, mode (whether or not a song is in a major or minor minor key), and time signature. Variables that should not necessarily have any impact on a song’s popularity were not included, such as latitude or longitude, artist ID, title, etc.

Artist popularity and familiarity were not taken into consideration, since Blackbird Records will be producing music for new and freshly-started musical artists, not ones who have already been signed. Subjective categories like “danceability” and “energy”, both measured from listeners’ viewpoints, were not included in the linear regression.



*Figure 6.* Displaying the results of the linear model.

The significance codes show that, because of how much smaller their p-values are compared to 0.05, the tempo is significant to $song.hotttnesss, as is loudness. Loudness is the most significant variable, with three stars in comparison to duration. The other variables do not have p-values under 0.05, showing that they are, in the end, not significant to song hotttnesss.

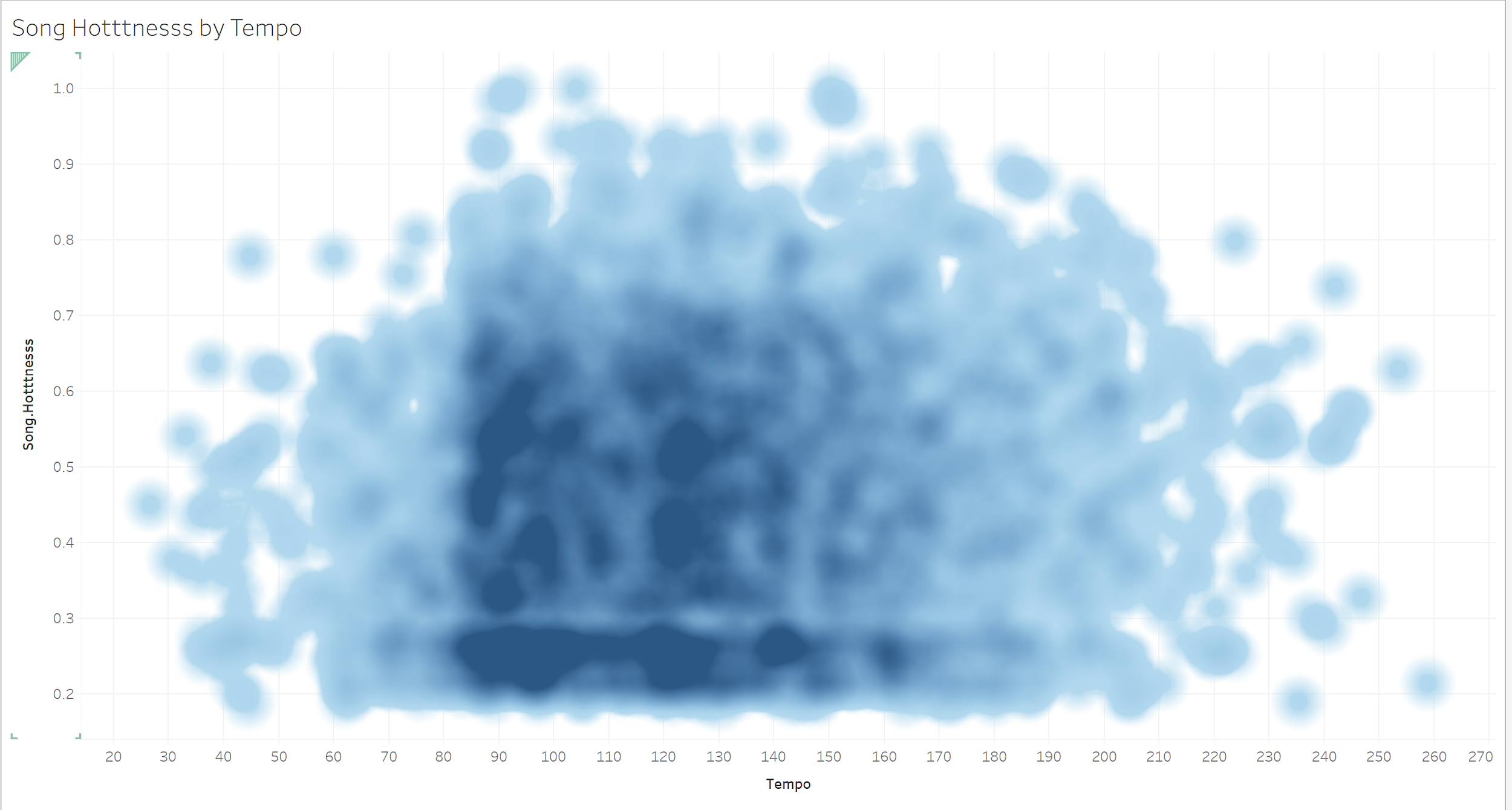
The coefficient for duration is positive but extremely small, showing that in a linear model, an increase in duration would marginally increase the popularity of the song. The same is the case for the variable loudness, although again, the coefficient is very tiny.

**Github Link to Project**

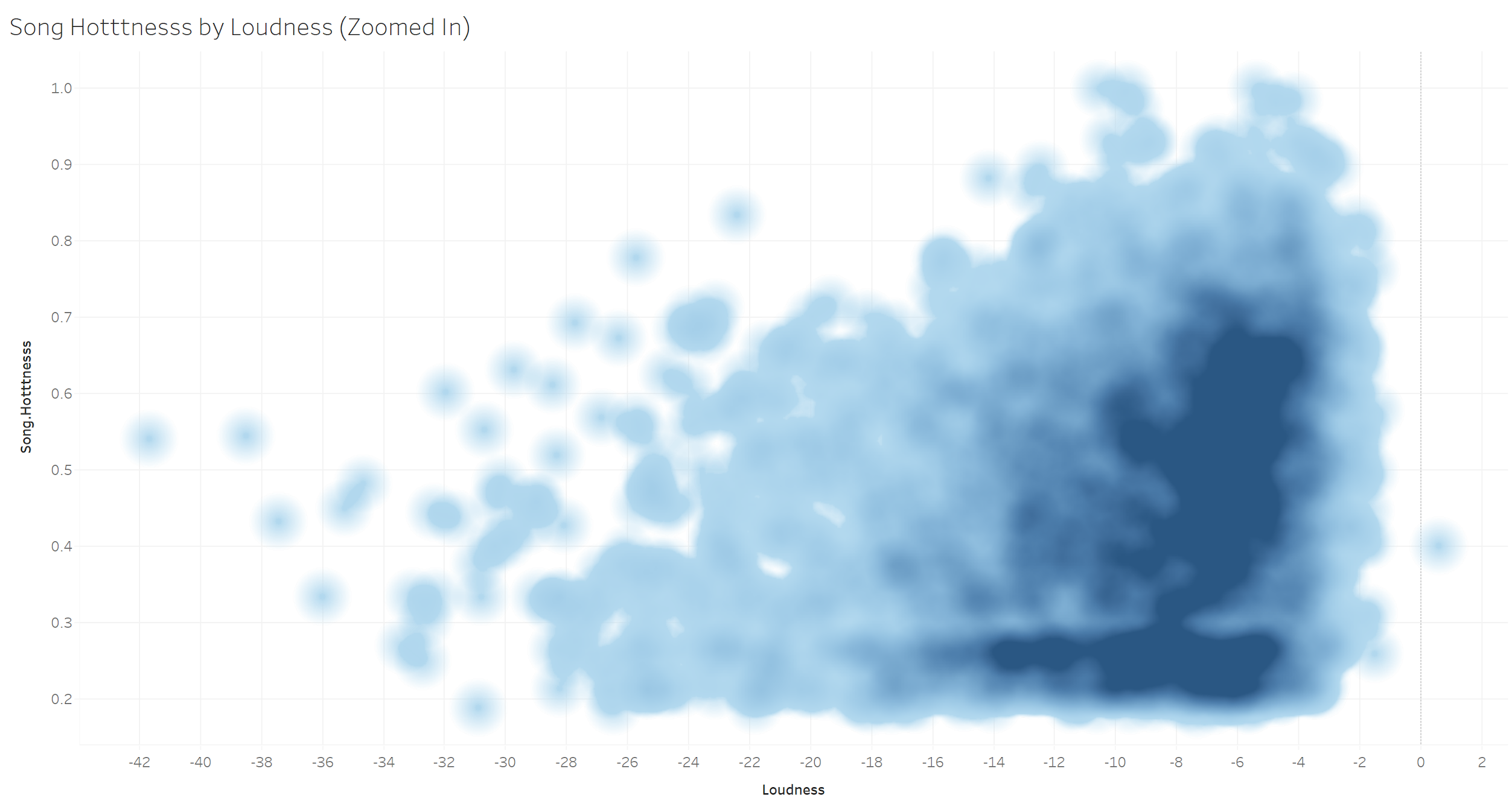
<https://github.com/kristaoneill07/csuglobalmis500>

**Visualizations**

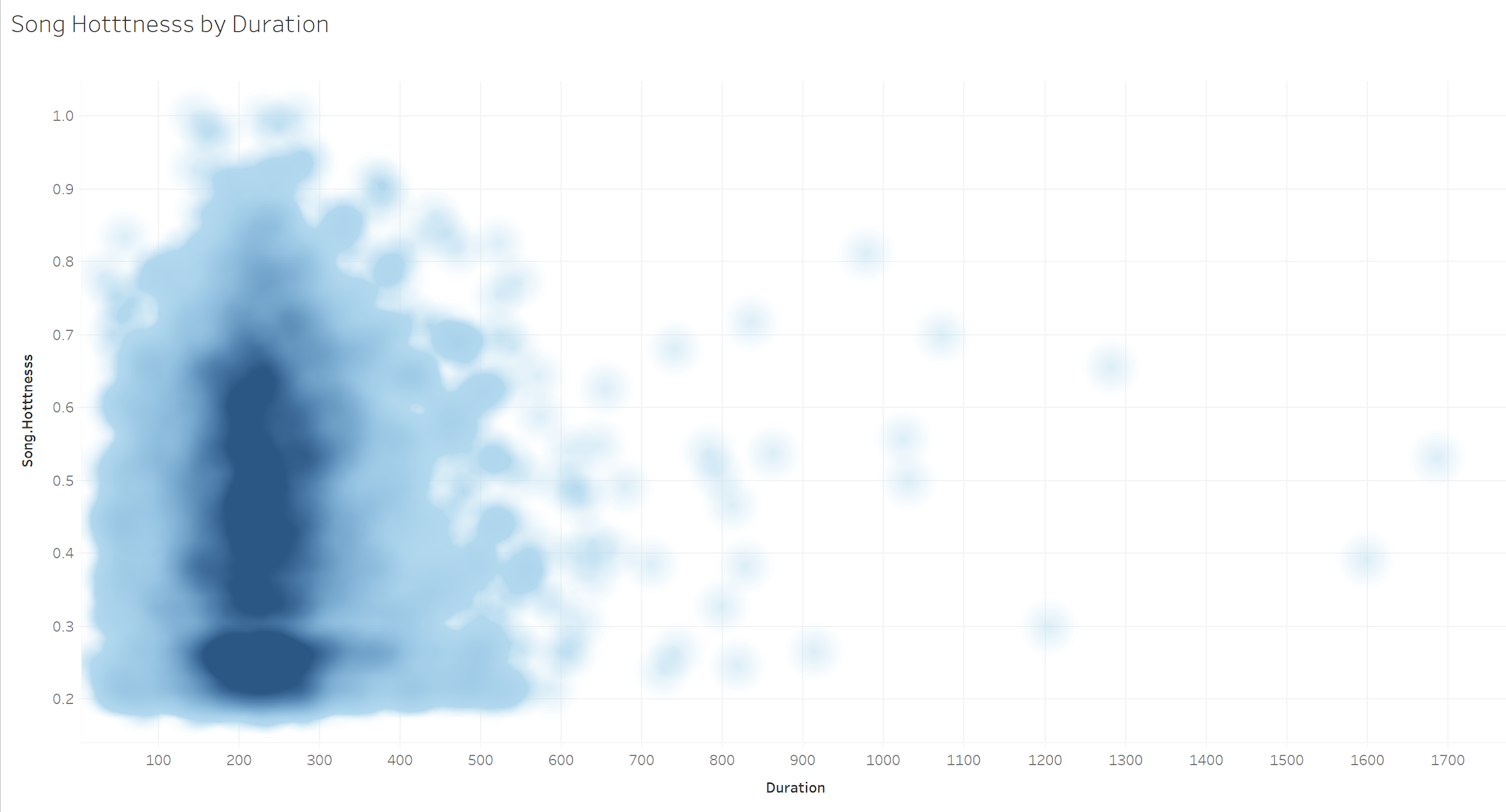
The cleaned dataset was exported to a .csv file for visualization in Tableau.



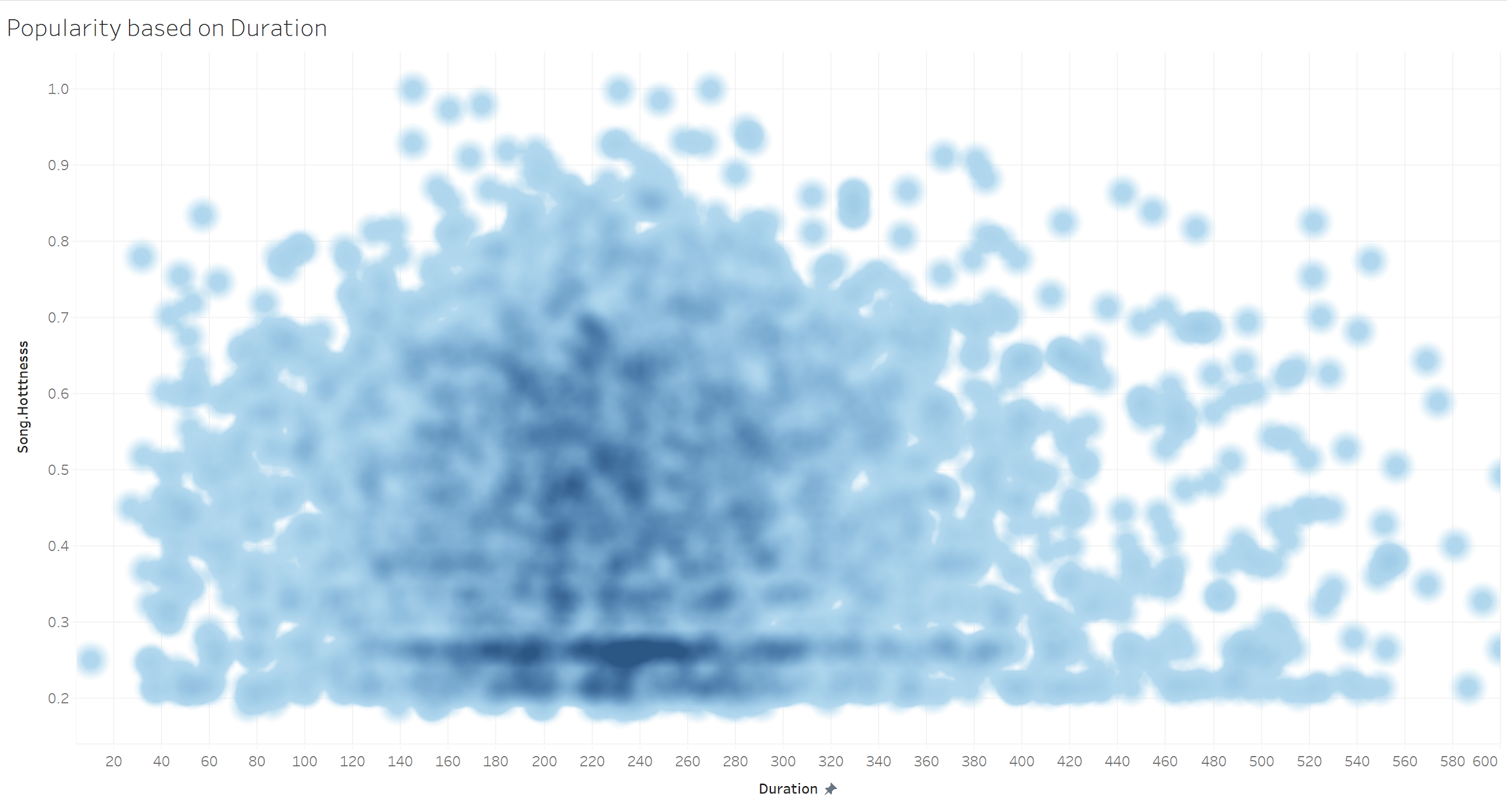
*Figure 7.* A graph of song hotttnesss based on tempo.



*Figure 8.* A graph of song hotttnesss based on volume. Songs seem to mostly fall in the volume range of -12 to -4, while song popularity appears to increase as volume increases.



*Figure 9*. A graph of song hotttnesss by duration.



*Figure 10.* A graph of song hotttnesss by duration, range 0-600 seconds only.

**Discoveries**

The previous t-test showed that tempo was significant to the popularity of a song. However, the samples used in that t-test were for values greater than 145.11, and less than 97.03 bpm. The visualization created in Tableau suggests that the majority of songs with high hotttnesss scores actually fall in the bpm values of 84 to 150.

It is interesting to see that the density maps for tempo show that there seem to be many songs within two separate tempos, around 90 bpm and 120 bpm. Additionally, there is a sharp increase in song hotttnesss between the values of roughly 79 bpm and 82 bpm.

In both visualizations, it can be seen that a simple linear model does not really fit the dataset, nor represent the relationship of tempo to song popularity. A quadratic model would possibly be a better fit.

It was surprising to find that some songs extended up to 1700 seconds long, and still received a hotttnesss score of above 0.5, or 50%. This could be an indicator that a record label should be willing to take a chance on oddball songs, as they could have equal success as songs that fit the average formula.

Finally, Blackbird Records’s original inquiry focused only on tempo and duration. However, the variable of greatest significance turned out to be loudness. The visualization of popularity based on loudness (see Figure 8) shows that it would benefit Blackbird Records to pay close attention the volume at which they mix their songs.

**Conclusion**

Blackbird Records now has an idea of what factors are the most important when it comes to profitably mixing songs. When the variables of preexisting artist familiarity and fame are removed, along with other subjective items, the factors tempo and loudness are, indeed, key deciders in how popular the song will become. However, song duration is not a significant variable.

As they move forward, they should pay close attention to the volume of the songs; ideally songs produced at the -11 dB to -4 dB range will have the most success. They should also aim to produce songs with a tempo between 85 and 150 bpm.

References

Billboard (2020). The hot 100: Week of May 2, 2020. [https://www.billboard.com/charts/hot-100/2020-05-02](https://www.billboard.com/charts/hot-100/2020-05-02?rank=2)

Dobrota, S., and Ercegovac, I. (2015). The relationship between music preferences of different mode and tempo and personality traits – implications for music pedagogy. *Music Education Research 17(2), 234-247*. <http://eds.b.ebscohost.com.csuglobal.idm.oclc.org/eds/pdfviewer/pdfviewer?vid=2&sid=ac438a55-ab6d-4999-af2e-443d1ebd9ff9%40sessionmgr103>

Kabacoff, R. (2017). Subsetting data. *Quick-R by DataCamp.* <https://www.statmethods.net/management/subset.html>

Kassambara, A. (2018). Regression analysis: Multiple linear regression in R. *Statistical tools for high-throughput data analysis.* <http://www.sthda.com/english/articles/40-regression-analysis/168-multiple-linear-regression-in-r/>

Liu, Y., Liu, G., Wei, D., Li, Q., Yuan, G., Wu, S., Wang, G., & Zhao, X. (2018). Effects of musical tempo on musicians’ and non-musicians’ emotional experience when listening to music. *Frontiers in Psychology*. <https://doi.org/10.3389/fpsyg.2018.02118>

Million Song Dataset (n.d). An example track description. <http://millionsongdataset.com/pages/example-track-description/>

Nasreldin, M. (2018). Song popularity predictor. *Towards Data Science.* <https://towardsdatascience.com/song-popularity-predictor-1ef69735e380>

Vatshayan (2018). Song-classification. *Github*. <https://github.com/Vatshayan/Song-Classification/blob/master/music.csv>

Warton, D., Lyons, M., Stoklosa, J., & Ives, A. (2016). Three points to consider when choosing a LM or GLM test for count data. *British Ecological Society*. <https://doi.org/10.1111/2041-210X.12552>

Xue, A., and Dupoux, N. (2014). Predicting a song’s commercial success based on lyrics and other metrics. *Stanford University*. <http://cs229.stanford.edu/proj2014/Angela%20Xue,%20Nick%20Dupoux,%20Predicting%20the%20Commercial%20Success%20of%20Songs%20Based%20on%20Lyrics%20and%20Other%20Metrics.pdf>

Yau, C. (2020). Significance test for MLR. *R Tutorial*. <http://www.r-tutor.com/elementary-statistics/multiple-linear-regression/significance-test-mlr>