**1. Introduction**

The “Music Genre” dataset is from the Kaggle dataset, it is used to train a machine learning model to predict what is the genre of an input music track based on several attributes, for example its popularity, key, energy, tempo, liveness, etc. The dataset includes detailed (both numerical and categorical) attributes for each music track and their genre (a categorical variable) to train a module. The detail definition of variables and data types are shown in Table 1. We choose this dataset because lots of measurement attributes of a given music track have already been transferred to numerical variable according to the Spotify algorithm, so it is easier to build a regression model and do more analysis among these variables.

The goal of the project is to utilize the knowledge and methods we learned in this course and answer our two research questions by cleaning our raw dataset, creating dummy variables, performing EDA analysis, creating correlation plot, running anova test, and constructing a GLM and logistic regression model.

The following are two research questions proposed based on the goals and the solution methods to be used:

* Does the popularity of music vary greatly depending on the genre of their music? There are 11 different music genres in this dataset, such as blues, country, jazz, electronics, rap, rock, etc. We would like to see whether the popularity means of blues, jazz, and electronics are significantly different. A One-Way ANOVA test will be used to solve this problem.
* Based on the characteristic value of the music, can we predict if this music contains certain music features? For example, we would like to determine the accoustiness of music based on other numerical music features: Tempo, loudness, liveness, and etc.

**2. Methods**

To address the first question, we performed a One-Way ANOVA test. ANOVA can be used to test the equality of three or more population means by analyzing sample variances to determine whether there is a relationship between them. This question has only one independent variable - music genre, so we will use the One-Way ANOVA test here. We do not need other methods as One-Way ANOVA test is the best option. ANOVA is more useful than a two-sample t-test because it is less likely to submit a Type I error. Also, the four assumptions for the ANOVA comparing three or more means are satisfied here:

* The popularity of the three music genres are approximately normally distributed and there are 5,000 records for each music genre, so we can assume that they are all normal distributed. The distribution results are shown in Figure 20 to Figure 22.
* The popularity of the three music genres are independent of each other.
* The variances of the three music genres are equal. Bartlett's test showed that the variances of the three groups were not significantly different (p=0.7831). The result is shown in Figure 23.
* The samples are simple random samples, one from each of the populations.

So based on the distribution plot and Bartlett's test, the data seem to fit well with an ANOVA model. These methods in turn increase our confidence in the results obtained.

If there is a significant difference in the popularity means, we can use TurkeyTest to do further research to examine which genres popularity differ significantly.

For the second question, we build Generalized Linear Model to predict the music acousticness based on the other valid numerical music features. We also take the music genre into account by creating the dummy variables. We use the backwards selection method which starts with all predictors added to the model and removes the predictors having the least impact on the model coefficients. AIC stands for Akaike information criterion, the lower the AIC the lower the degree of information loss. So it’s a great indicator to see the validity of this model.

**3. Analysis**

After the data cleaning process, there are 50,000 observations and 28 variables in the dataset. The summary and structure of the cleaned dataset are shown in Table 2 and Table 3.

The distributions of the independent variables are shown in Figures 10 to 19. We can see that the distributions of "danceablity", "tempo" and "duration\_ms" are close to normal, the distributions of "energy" and "loudness" are left skewed, and the others are right skewed.

To solve the first research question, we would like to see the relationship between music genre and popularity, so we created a music genre and popularity table (Table 4). Additionally, we created bar (Figure 4)and box plots (Figure 5 & Figure 6) to visualize music genre popularity rankings and the popularity distribution of each music genre.

From Table 4 and Figure 4, we can see that there are 10 music genres, each with 5,000 counts. However, their average popularity is quite different. The top three music genres are Rap, Rock and Hip-Hop. Rap is the most popular genre of music, with an average popularity of 60.50, while Anime is the least popular, with an average popularity of only 24.27.

Figure 5 shows the distribution of popularity by music genre. From Figure 5, it can be seen that the popularity distribution of Hip-Hop, Alternative, Country, Jazz, and Classical is close to a normal distribution, while the popularity distribution of Rap, Rock, Electronic, Blues, and Anime is right-skewed distribution. We can conclude that Rock and Rap are very popular and they tend to have high popularity scores.

To solve the second research question, we clean our dataset by keeping all of the relevant numerical values and identify their music genre by creating the dummy variables. For example, if the music is hip pop, then the value for hip-popness is 1 and vice versa. Then we use glm function and backward method to find the best fit model.

**Anova test for the first research question:**

As there are too many records(15,000) in the dataset of the three genres, we need to take a subset for the ANOVA test to ensure the accuracy.

* State the hypotheses and identify the claim

·Null Hypothesis

The null hypothesis will state that there is no difference between the means for the three populations.

H0: μ1 = μ2 = μ3

·Alternative Hypothesis

The alternative hypothesis will then state that at least one mean is different from the others. There is a difference in mean of the popularityl exists among Jazz, Blues, and Electronic.

H1: At least one mean is different from the others.

* Find the critical value

After running anova test in R, we found that d.f.N = 2, and d.f.D = 7,498. The P-value is less than 2e-16. The ANOVA test results are shown in Figure 24.

In the F-table shown in Figure 25, the columns correspond to the degrees of freedom of the numerator and the rows correspond to the degrees of freedom of the denominator. We can find the critical value using d.f.N and d.f.D. The critical value is 2.9957.

* Compute the test value

After running anova test in R, we found that the F test value is 244.9. The result is shown in Figure 24.

* Make the decision

Since the test value of 244.9 is greater than the critical value of 2.9957 we should reject the null hypothesis. Also, as the p-value is less than 2e-16, which is less than alpha value 0.05, we rejected the null hypothesis.

* Summarize the results and explain where the differences in the means are

Since we rejected the null hypothesis there is enough evidence to conclude that a difference in mean of the popularityl exists among Jazz, Blues, and Electronic.

While ANOVA's F-test for each music genre indicated that the three music genres had different popuarity, it did not say which music genre was different from the others. Multiple comparisons can solve this problem. The TukeyHSD() function provides a pairwise test for the difference in means between groups. Figure 26 shows that the p-values for jazz - blues, blues - electronic, and jazz - electronic are all equal to 0, which means the popularity of the three music genres are all different from each other. The pairwise comparison graph is shown in Figure 27. Treatments with a confidence interval containing 0 in the graph indicate that the difference is not significant (p>0.5). We can see that the popularity of the three music genres differs significantly. As can be seen from Figure 28, the average popularity rankings from high to low are: jazz, electronic, blues.

**GLM for the second research question:**

First, we prepare a dataset consists all of the numerical values. Acousticness, danceability, duration\_ms, energy, instrumentalness, liveness, loudness, speechiness, tempo, valence, and the dummy variables are Alternativenss, Animeness, Blueness, Classicalness, Countriness, Electronicness, Hiphopness, Jazzness.

We have the GLM for Acousticness as Acousticness= 0.8592-0.2099\*danceability-5.256\*10^-08\*duration\_ms-8.668e-01\*enerby+1.883e-02\*instrumentalness+4.727e-02\*liveness-1.460e-03\*loudness+1.154e-01\*speechiness-3.507e-04\*tempo+1.326e-01\*valence+6.638e-02\*Animeness+4.267e-02\*Blueness+1.946e-01\*Classicalness+4.169e-02\*Countriness-2.503e-02\*Hiphopness+1.145e-02\*Jazzness+-2.793e-02\*Rapness. All of the Pr values for these predictors are approximately 0. It means that they are make significant impact on the acousticness. The AIC value is -23144. It means that model has an extremely lower degree of information loss.

We also notice that the null deviance and the residual deviance. The null deviance indicates how well the response variable can be predicted by a model with only an intercept term. The residual deviance tells us how well the response variable can be predicted by a model with p predictor variables. The lower the value, the better the model is able to predict the value of the response variable. We can perform a Chi-Square statistic test. The lower the p-value, the better the model is able to fit the dataset.

Chi-Square=Null deviance-Residual deviance

So for our model, the Chi-Square=5825.5-1841.4=3984.1. The predictor variables degrees of freedom is 49999-49983=16. The P value is at 0. Since this p-value is much less than0.05, we can conclude that the model is highly effective for predicting acousticness of the music.

**4. Interpretation & Conclusions**

According to the analysis of the first research question, we have confidence to conclude that the popularity of the three music genres differs significantly when alpha value is 0.05. The average popularity rankings from high to low are: jazz, electronic, blues. We can infer that jazz is more popular than the other two genres. From the model, we can see that hip pop and rap isn’t titled to acousticness. For further research, we can study which music metrics contribute the most to popularity.

**5. Tables and figures**

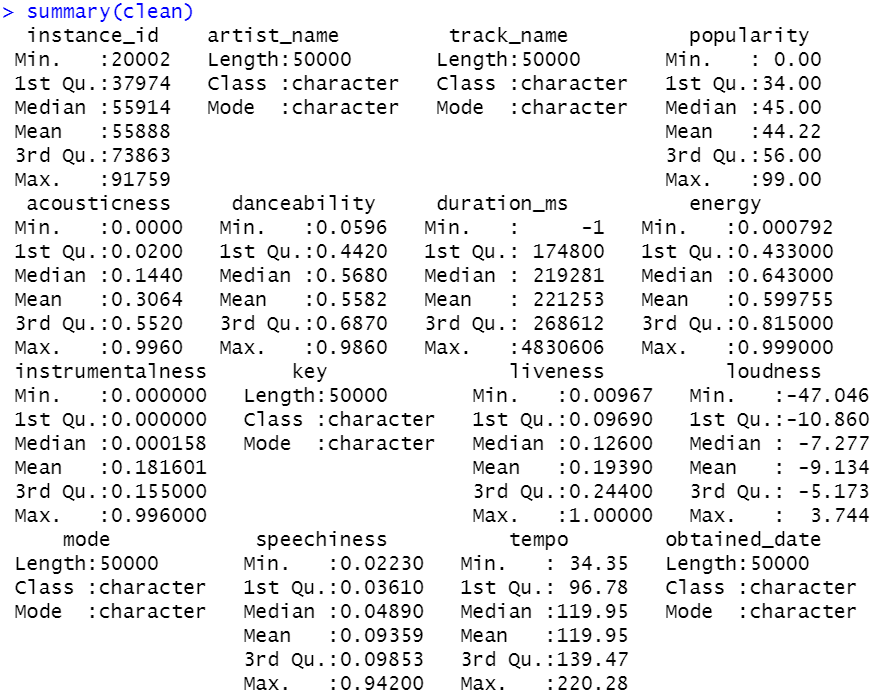
Table 1 shows tested parameters and their definitions. As shown in the table, each attribute in the dataset is interpreted. So you can have a better understanding on the chosen music.

***Table. 1. Characteristic for the selected music***

|  |  |
| --- | --- |
| **Characteristics** | |
| **Acousticness** | A confidence measure from 0 to 1 indicate how acoustic the music is. 1.0 represents high confidence the track is acoustic. |
| **Danceability** | Ranges from 0 to 1. It tells how danceable the music is. |
| **Duration\_ms** | It tells how long the music is. |
| **Energy** | It measure from 0 to 1 and represent the intensity and activity of the music. |
| **Instrumentalness** | It indicates amount of vocal content in the music. It ranges from 0 to 1. |
| **Key** | The pitch in the music. |
| **Liveness** | It indicates the presence of any audience in the recording. |
| **Loudness** | The greater the value is, the louder the music is. |
| **Mode** | It indicates either major or minor. |
| **Speechiness** | Value above 0.66 describe the music are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech. Value below 0.33 most likely represent music and other non-speech-like tracks. |
| **Tempo** | The overall estimated tempo in beats per minute. |
| **Obtained date** | The data when it released. |
| **Valence** | It measure the musical positiveness of the music. The higher the value, the more positive the music is. |
| **Music genre** | The type of the music. |

Table 2 shows the summary of the cleaned dataset. Here, we can quickly check the mean, median, and range of each attribute.

***Table. 2. Summary of the cleaned dataset***



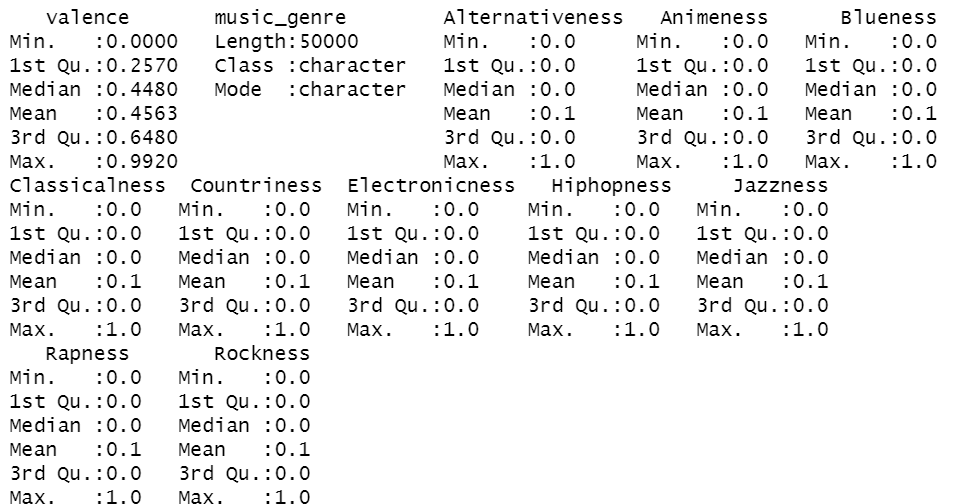


Table 3 shows the summary of the cleaned dataset. Here, we can quickly check the mean, median, and range of each attribute.

***Table. 3. Summary of the cleaned dataset***

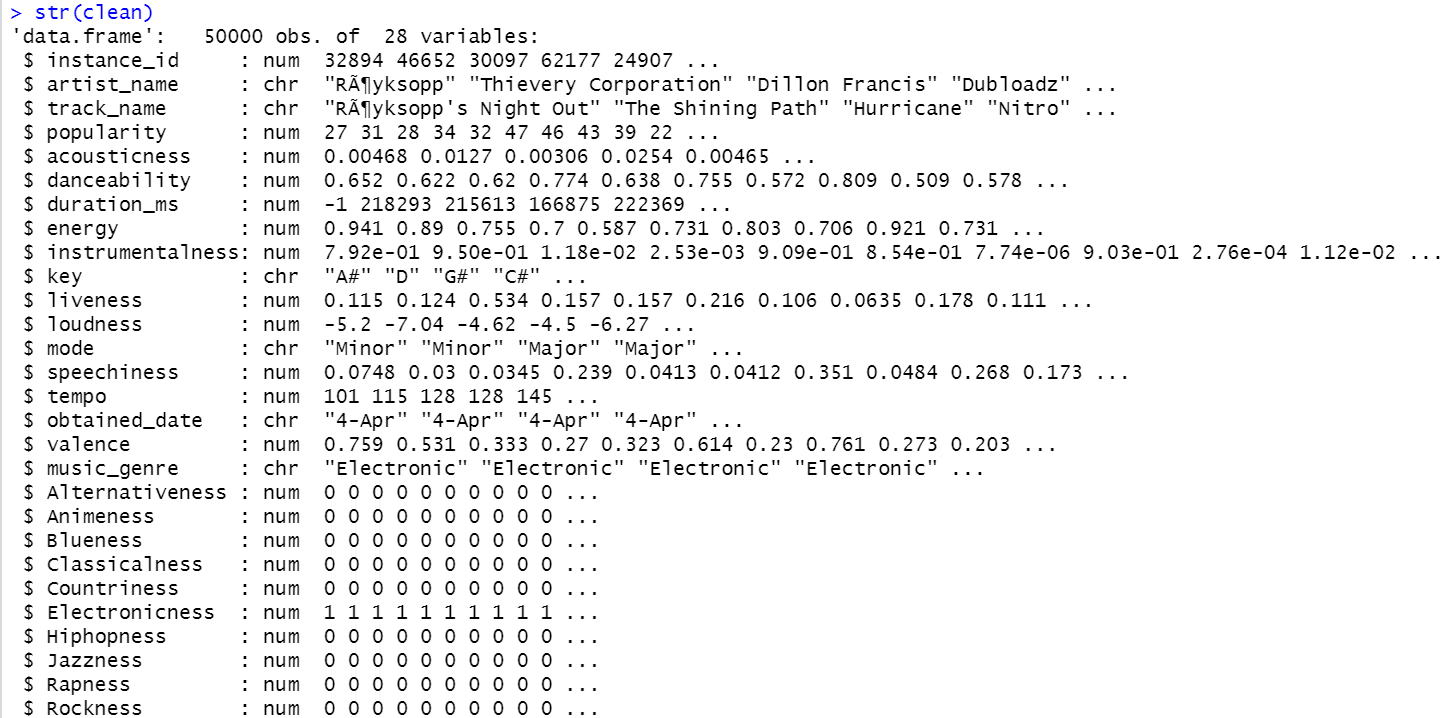
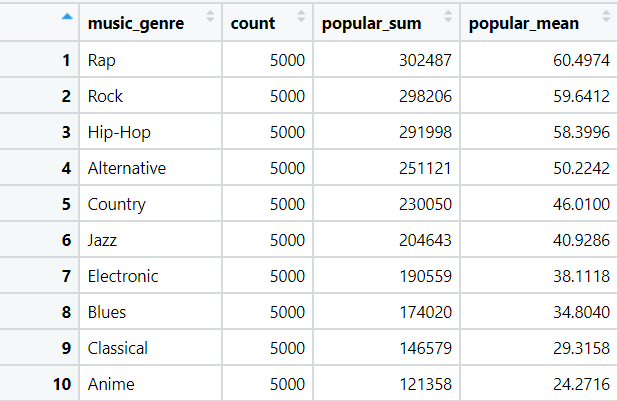
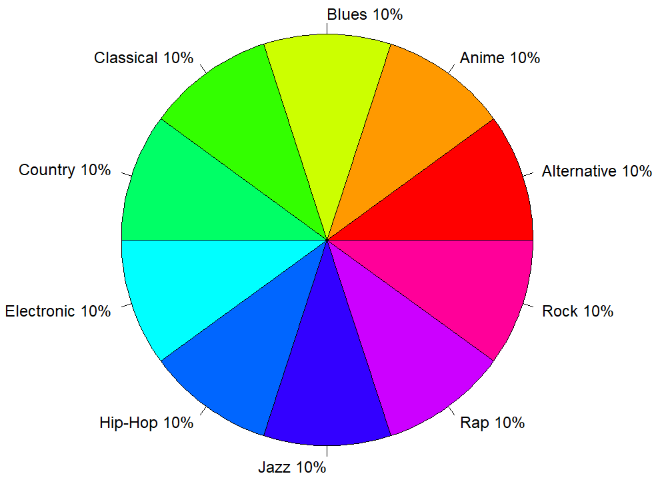


Table 4 shows the counts for each music genre and the related sums of popularity and averages of popularity in descending order.

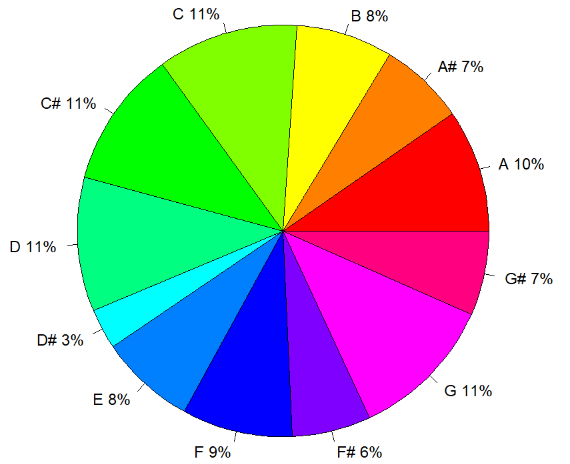
***Table. 4. Popularity & Music Genre***





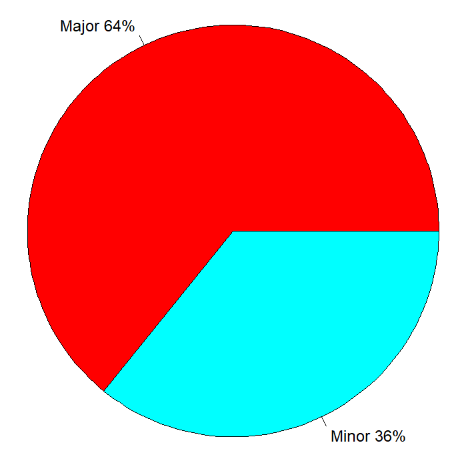
***Figure 1: Music genre distribution.***

Figure 1 shows the distribution of music genre. We can see that there are 10 music genre in total and are evenly distributed in our dataset.



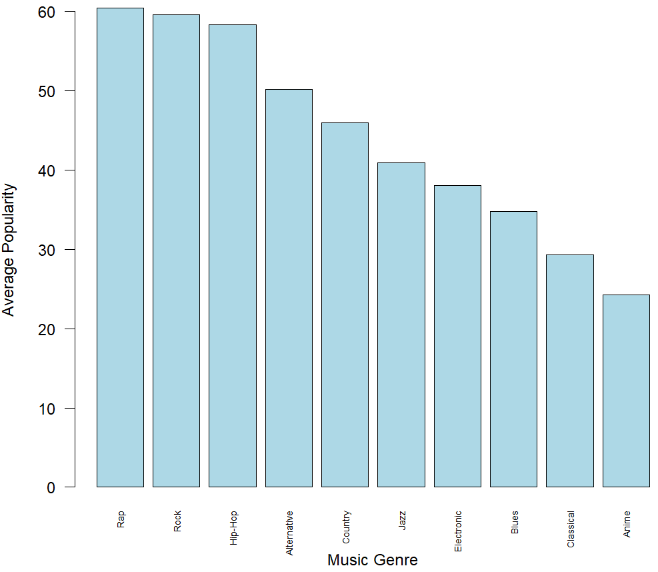
***Figure 2: Music key distribution.***

Figure 2 shows the distribution of music key. We can see that there are 13 music keys in total in our dataset. We know that there are 12 major keys and 12 minor keys which make up 24 keys all together. Our dataset coverd a little bit over 50% of all music keys.



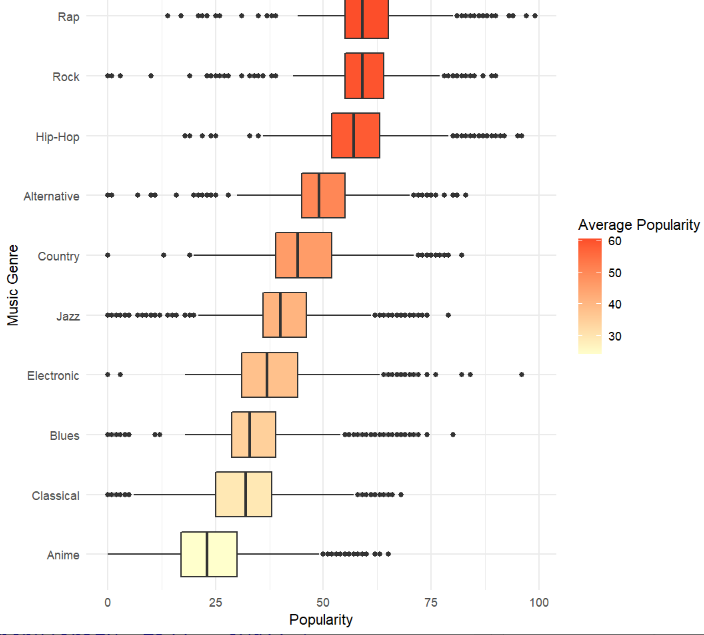
***Figure 3: Music mode distribution.***

Figure 3 shows the distribution of music mode. We can see that there are 2 music modes in our dataset. Most of the music in our dataset are major.



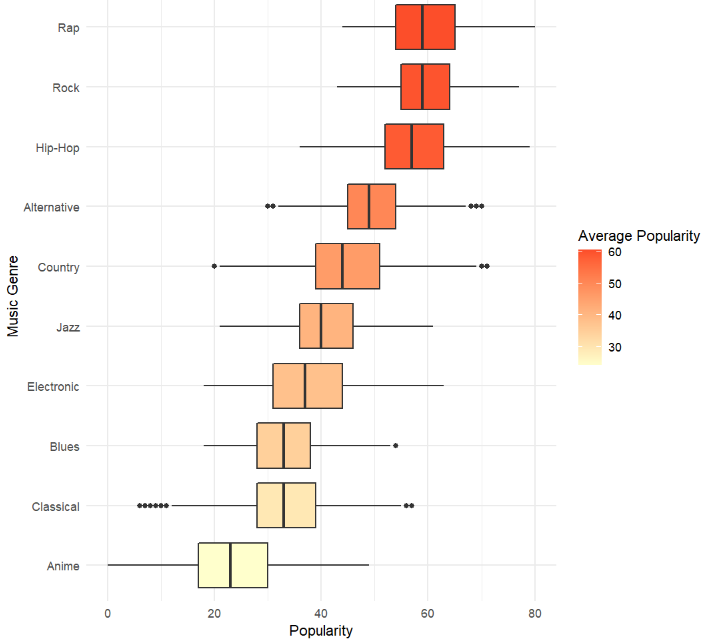
***Figure 4: Average popularity by music genre***

Figure 4 shows the average popularity for each music genre in descendin order. The top 3 music genres are Rap, Rock, and Hip-Hop.



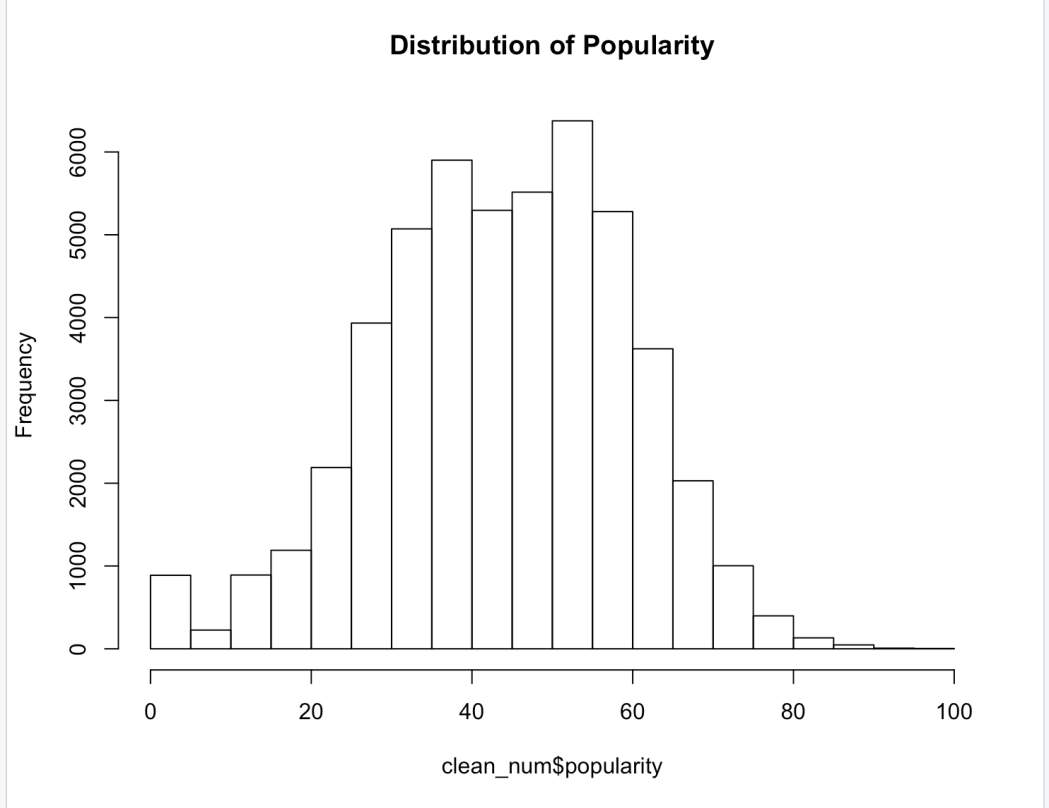
***Figure 5: Distribution of popularity by music genre***

Figure 5 shows the popularity distribution for each music genre (with outliers).



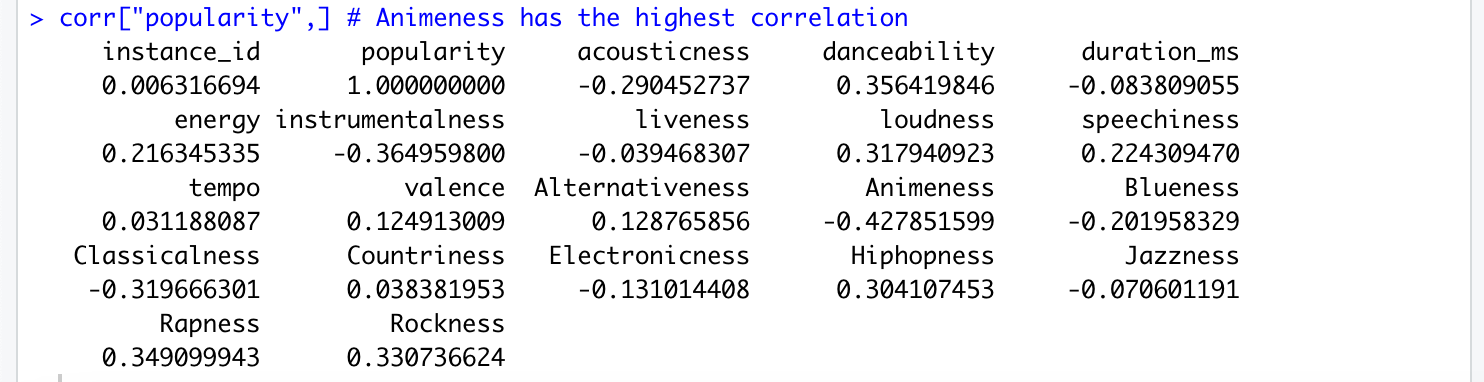
***Figure 6: Distribution of popularity by music genre (outliers removed)***

Figure 6 shows the popularity distribution for each music genre (without outliers).



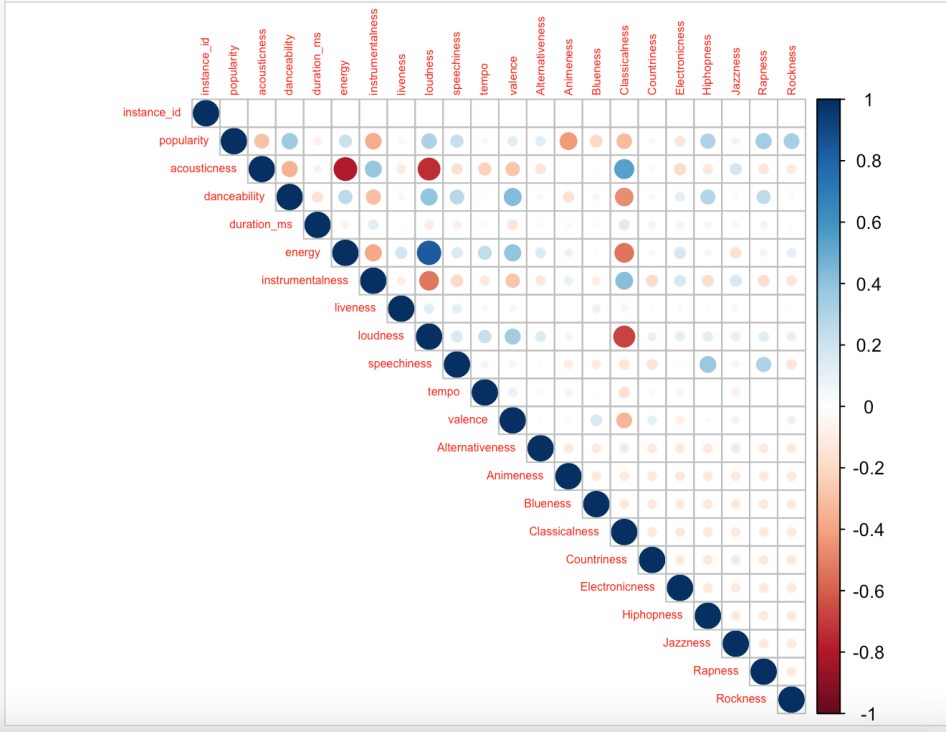
***Figure 7: Distribution of Popularity***

Figure 7 shows the distribution of the dependent variable - popularity in the music genre dataset.



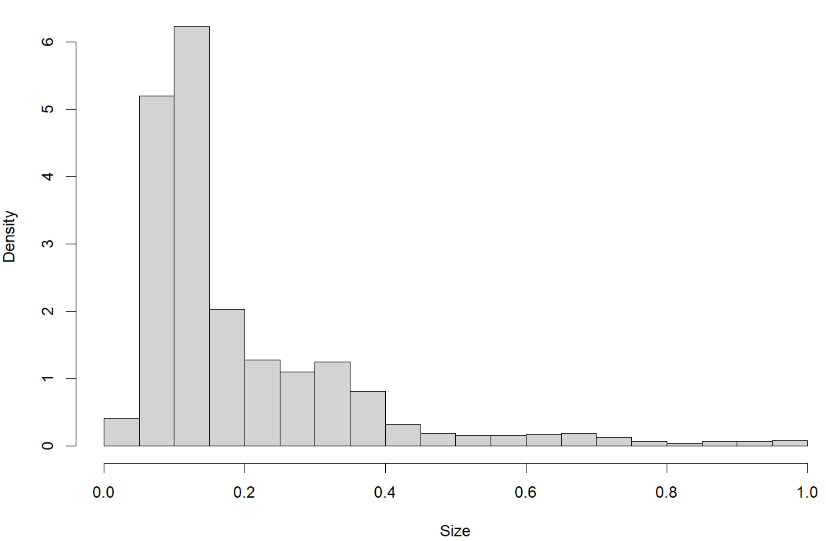
***Figure 8: Correlation matrix among “popularity” and other attributes***

Fiure 8 shows the correlatio matrix among dependent variable “popularity” and other potential independent variables.



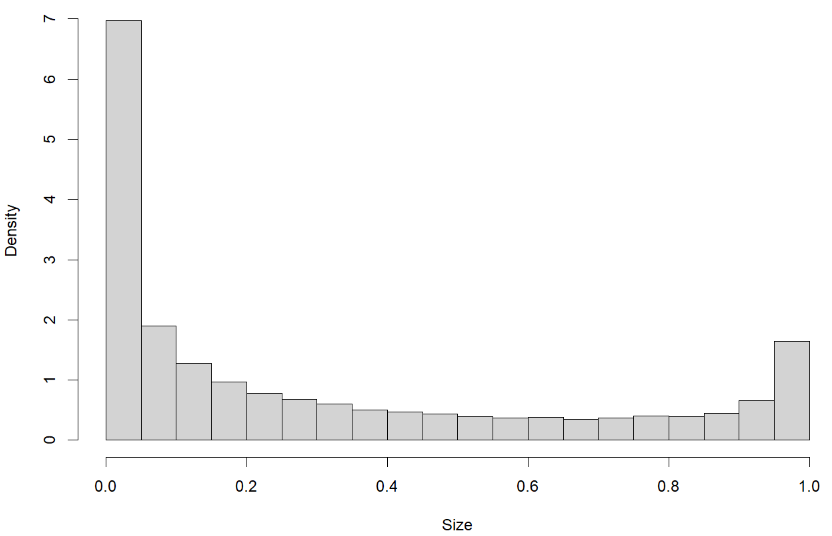
***Figure 9: Correlation plot of all numerical variables in music genre dataset***

Figure 9 shows a correlation plot of all numerical variables in the dataset. In this plot, the darker blue or red a point is, the stronger correlation between two variables.



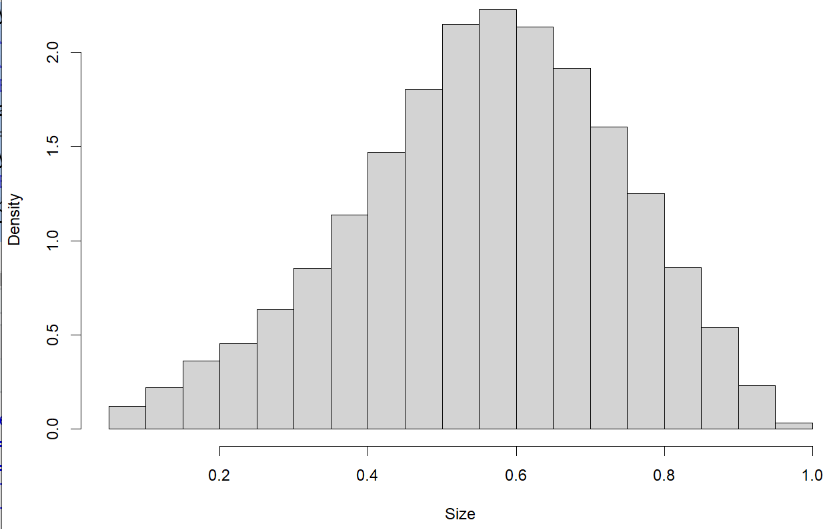
***Figure 10: Distribution of Liveness***

Figure 10 shows the distribution of the independent variable - liveness in the music genre dataset.



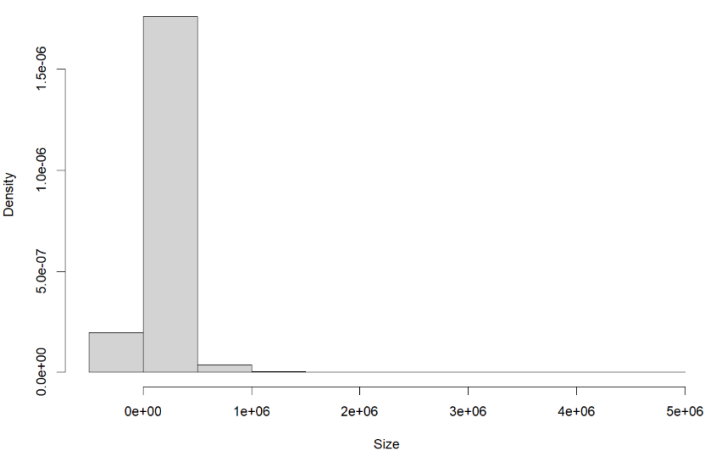
***Figure 11: Distribution of Acousticness***

Figure 11 shows the distribution of the independent variable - acousticness in the music genre dataset.



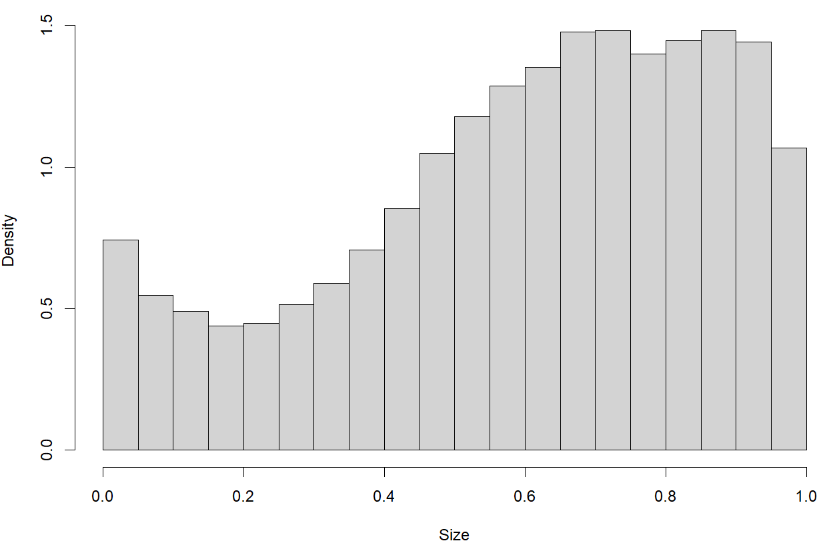
***Figure 12: Distribution of Danceability***

Figure 12 shows the distribution of the independent variable - danceability in the music genre dataset.



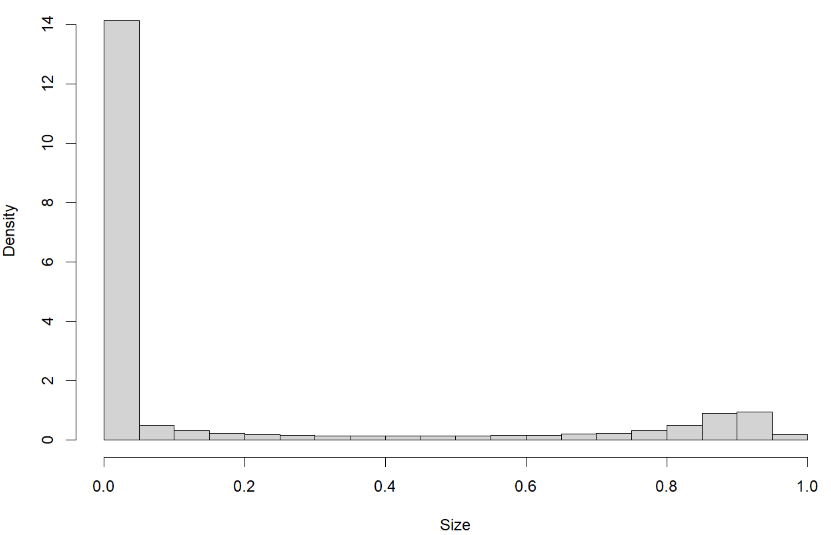
***Figure 13: Distribution of Duration\_ms***

Figure 13 shows the distribution of the independent variable - duration\_ms in the music genre dataset.



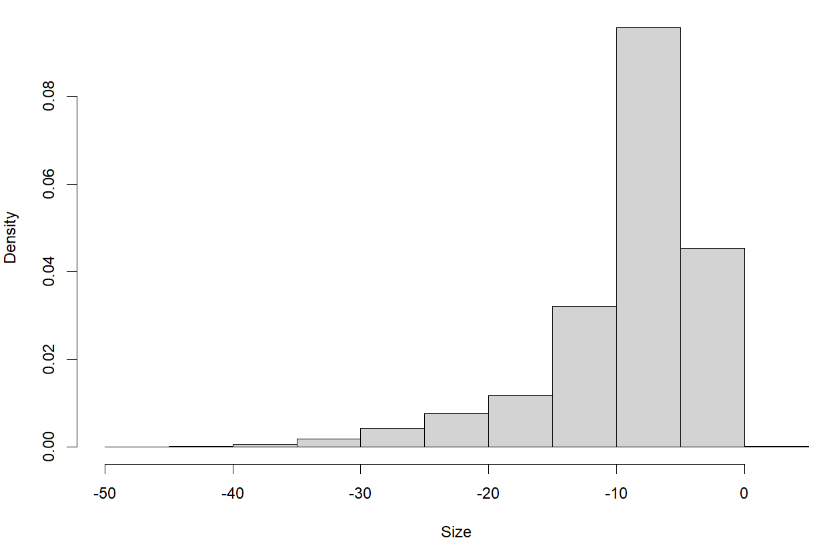
***Figure 14: Distribution of Energy***

Figure 14 shows the distribution of the independent variable - energy in the music genre dataset.



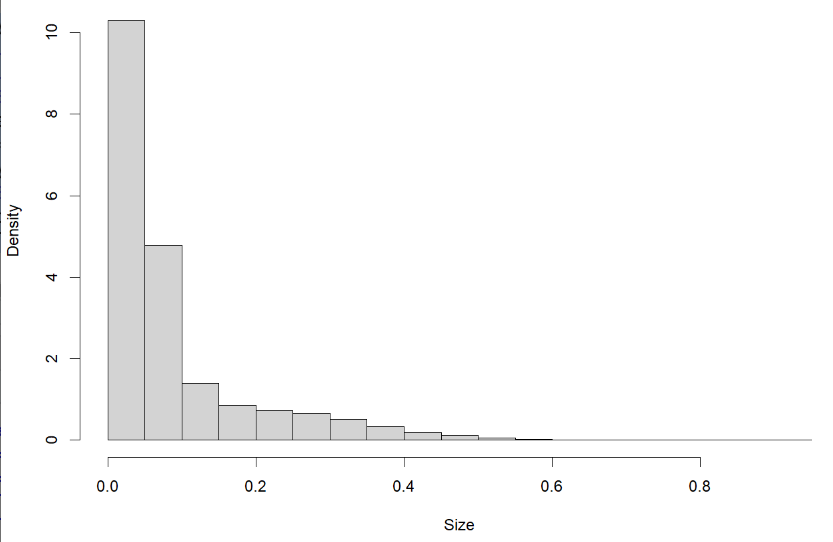
***Figure 15: Distribution of Instrumentalness***

Figure 15 shows the distribution of the independent variable - instrumentalness in the music genre dataset.



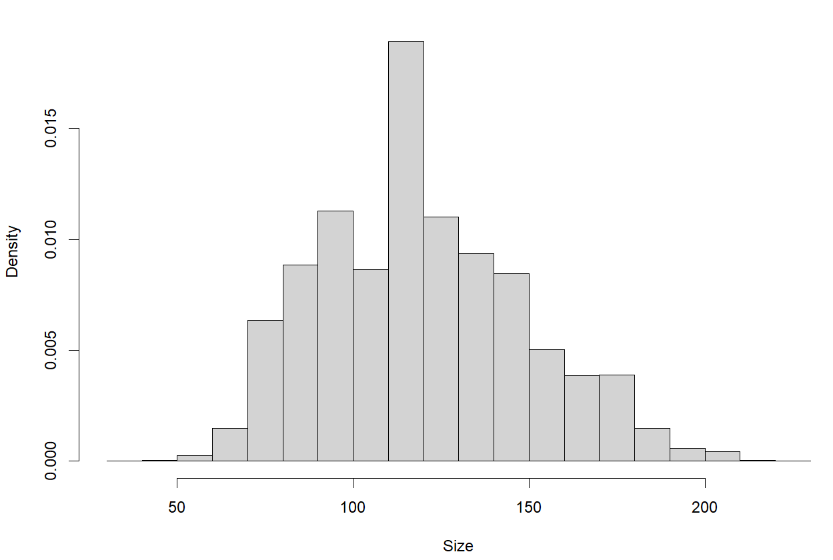
***Figure 16: Distribution of Loudness***

Figure 16 shows the distribution of the independent variable - loudness in the music genre dataset.



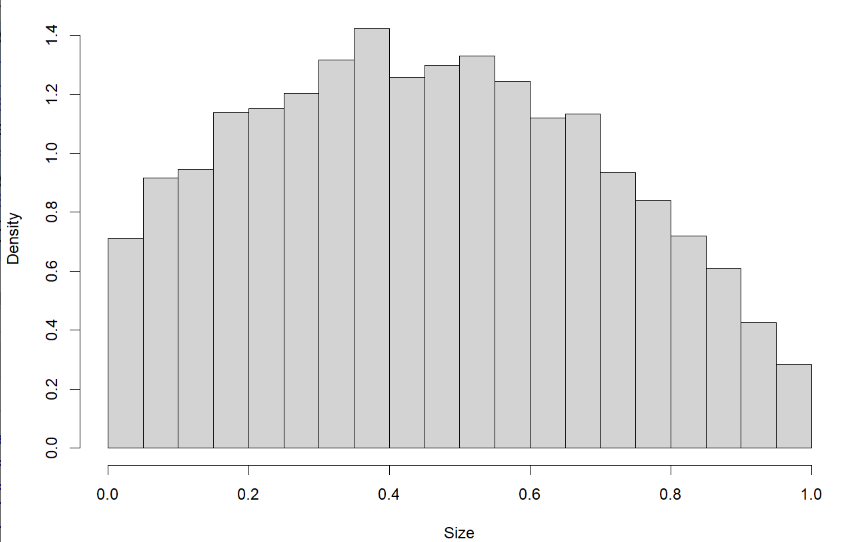
***Figure 17: Distribution of Speechiness***

Figure 17 shows the distribution of the independent variable - speechiness in the music genre dataset.



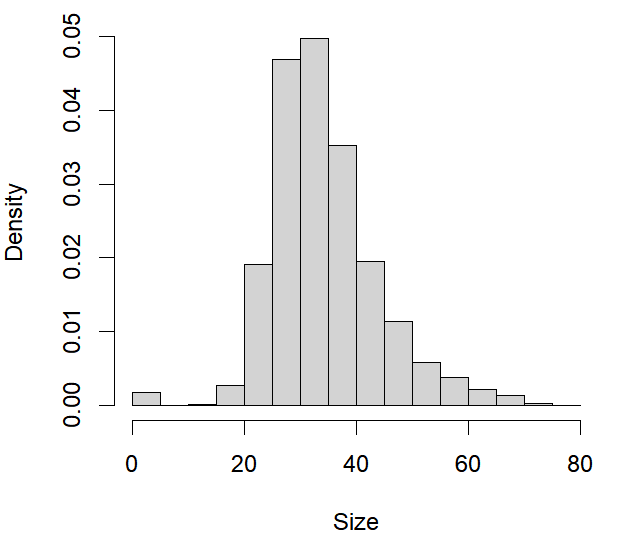
***Figure 18: Distribution of Tempo***

Figure 18 shows the distribution of the independent variable - tempo in the music genre dataset.



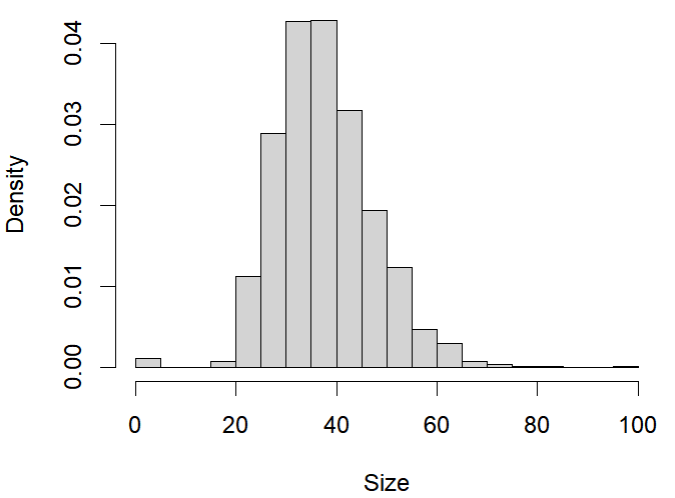
***Figure 19: Distribution of Valence***

Figure 19 shows the distribution of the independent variable - valence in the music genre dataset.

****

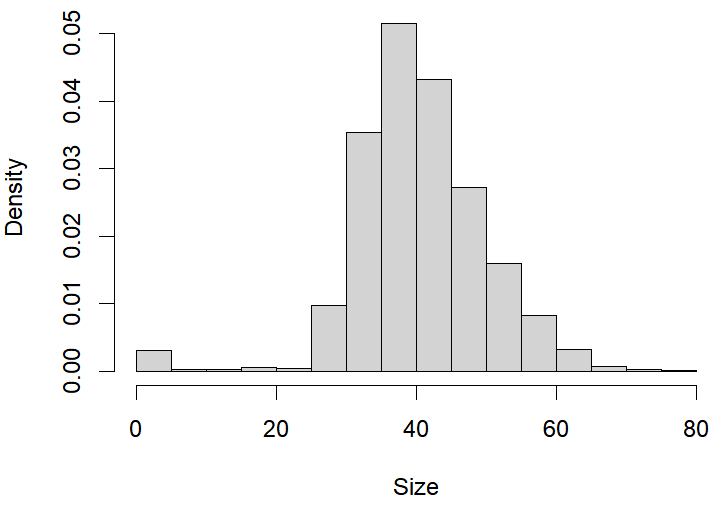
***Figure 20: Distribution of Blues Populatity***

Figure 20 shows the distribution of the popularity of Blues in the music genre dataset.

****

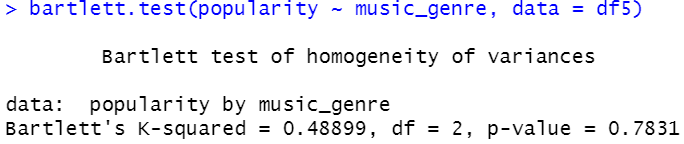
***Figure 21: Distribution of Electronic Populatity***

Figure 21 shows the distribution of the popularity of Electronic in the music genre dataset.

****

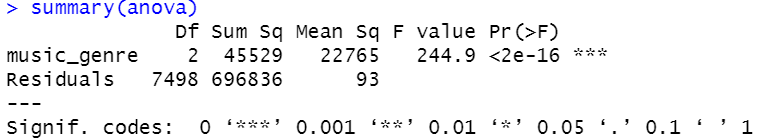
***Figure 22: Distribution of Jazz Populatity***

Figure 22 shows the distribution of the popularity of Jazz in the music genre dataset.

****

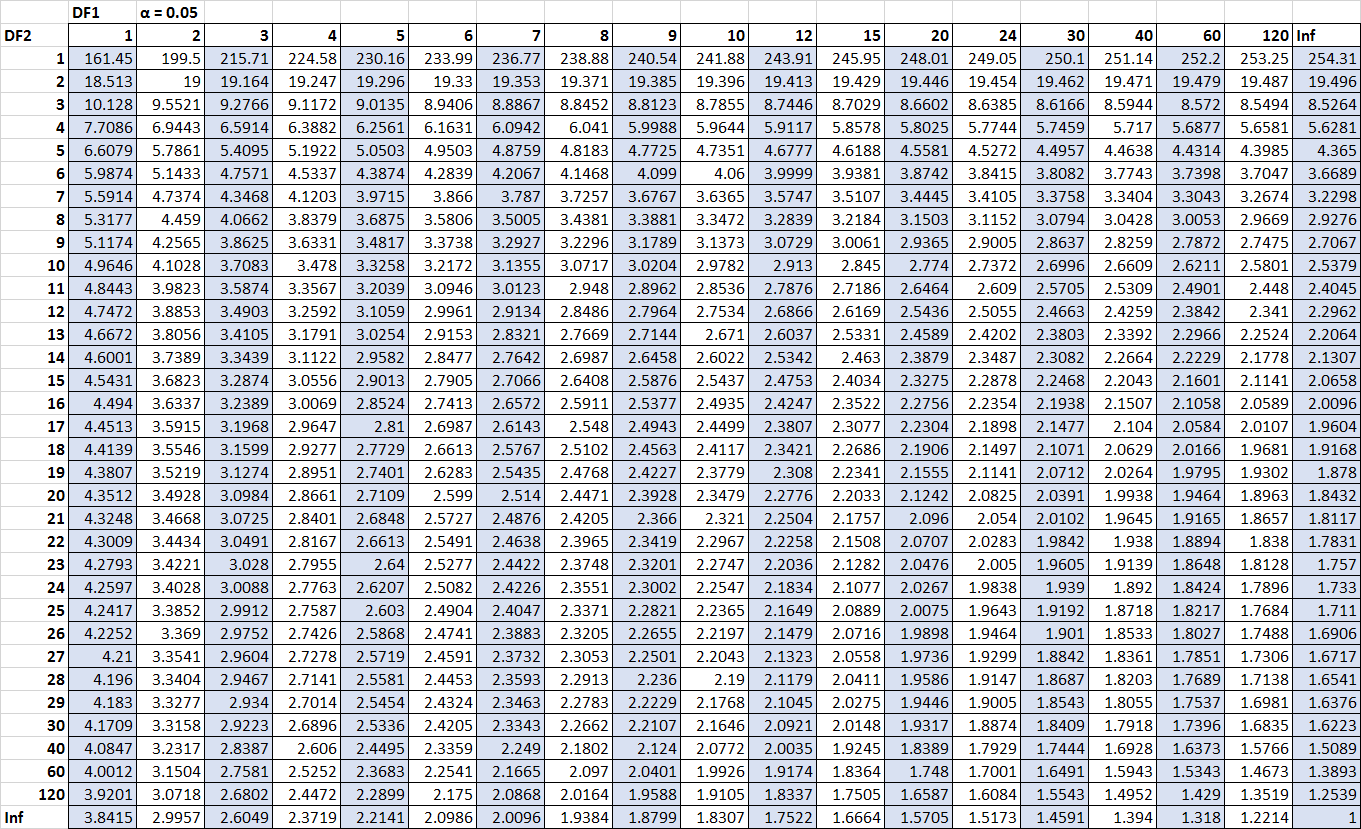
***Figure 23: Bartlett's test of the three groups***

Figure 23 shows the Bartlett's test of the three groups. Bartlett's test showed that the variances of the three groups were not significantly different (p=0.2431)

****

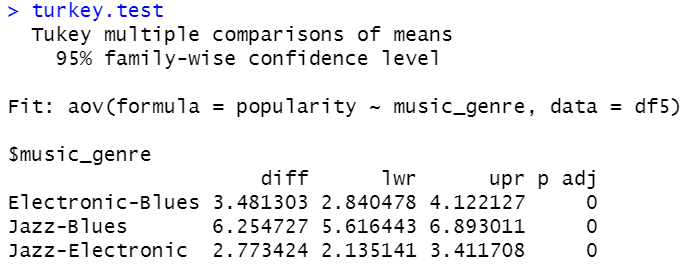
***Figure 24: Anova test result***

Figure 24 shows the Anova test result of the first research question.



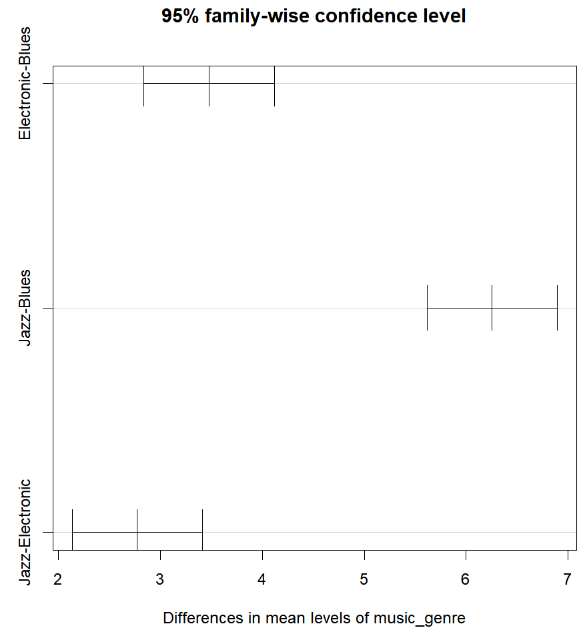
***Figure 25: F table (significance level = 0.05)***

Figure 25 shows the F table for significance level at 0.05.



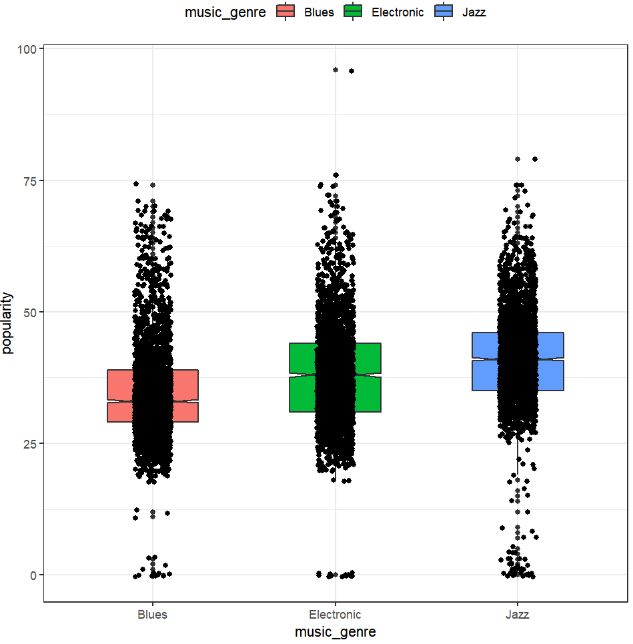
***Figure 26: Turkey test result***

Figure 26 shows the Turkey test result for the first research question



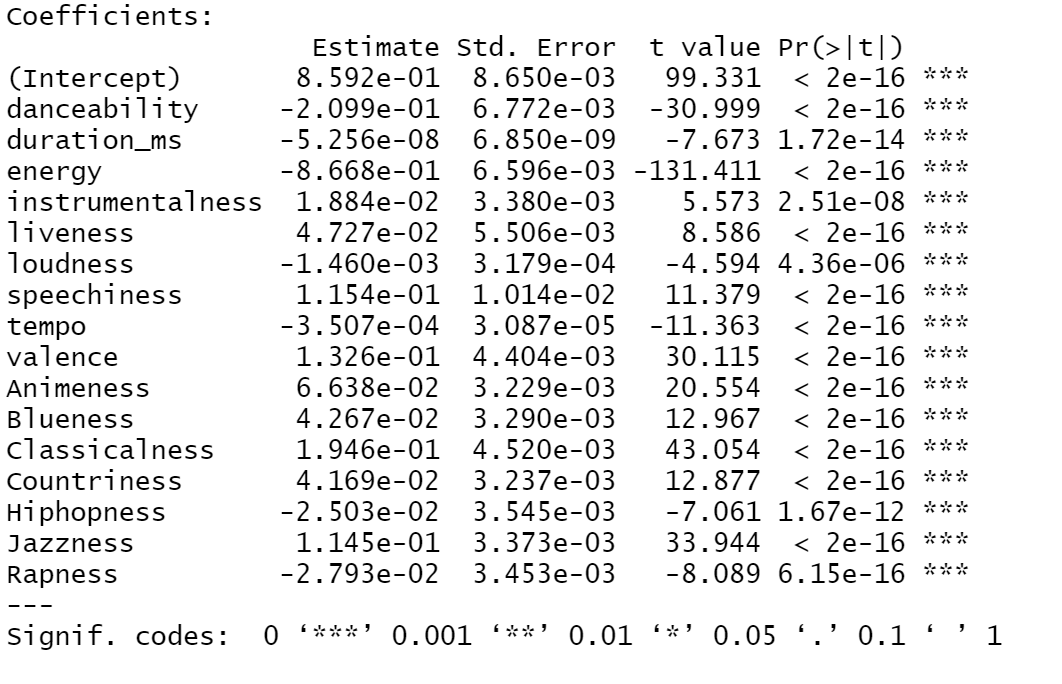
***Figure 27: Tukey HSD Mean Pairwise Comparison Plot***

Figure 27 shows the Tukey HSD mean pairwise comparison results.



***Figure 28: Boxplot of popularity for blues, jazz, and electronic***

Figure 28 shows the Boxplot of popularity for blues, jazz, and electronic. We can see the distributino comparisons among these three music genres.



***Figure. 29. Coefficients of the GLM***

**7. Reference:**

* Tiffany, K. (2018, February 5). *You can now play with Spotify's recommendation algorithm in your browser*. The Verge. Retrieved April 23, 2022, from https://www.theverge.com/tldr/2018/2/5/16974194/spotify-recommendation- algorithm-playlist-hack-nelson
* Bitten, H. (2019, June 5). Moneyball — Linear Regression. Medium. https://towardsdatascience.com/moneyball-linear-regression-76034259af5e
* Zach. (2020, October 21). How to Perform a Chi-Square Goodness of Fit Test in R. Statology. https://www.statology.org/chi-square-goodness-of-fit-test-in-r/
* Chi-square Goodness of Fit Test in R - Easy Guides - Wiki - STHDA. (n.d.). Www.sthda.com. <http://www.sthda.com/english/wiki/chi-square-goodness-of-fit-test-in-r>
* “5 Steps on How to Approach a New Data Science Problem.” Accessed April 20, 2022.<https://brainhub.eu/library/how-to-approach-data-science-problem/>.

‌