An Analysis of the Commercial Success of Popular Music

— Investigating Factors that affect the Sales of a song

1 Introduction

The digital music industry has seen a transformative shift in recent years, with live streaming and online platforms becoming the primary means for music consumption worldwide. This change has not only altered how music is distributed and consumed but also how its success, measured through sales and streams, is predicted. In this study, we aim to construct a predictive model for music sales, focusing on the impact of the different variables, using data from the "Top Songs of the World" dataset on Kaggle. Our research question. Our research question focuses on how streams, downloads, radio plays, and rating affect the prediction of music sales. The dataset utilized in this study is sourced from Kaggle, specifically from the "Top Songs of the World" dataset. This dataset compiled detailed information on songs that have achieved significant commercial success worldwide, including their sales figures, streaming counts, rates, and other pertinent attributes. The comprehensive nature of this dataset makes it an ideal foundation for analyzing the factors influencing music sales in the current digital age. We chose multiple regression analysis for its ability to elucidate the relationships between multiple variables and a continuous outcome. The paper is organized into sections that introduce the study, describe our methodology, present our findings, and discuss their implications in the broader context of the digital music industry, culminating in a conclusion that highlights key takeaways and suggests avenues for future research. This study offers insights into the dynamics of music popularity and sales, providing valuable information for stakeholders in the music industry.

2 Data Description

2.1 Summary Statistics

```
   Artist             Title                Year          Sales
Length:4850        Length:4850        Min.   :1901   Min.   : 4.149
Class :character   Class :character   1st Qu.:1967   1st Qu.: 5.069
Mode  :character   Mode  :character   Median :1981   Median : 6.438
                                      Mean   :1980   Mean   : 8.155
                                      3rd Qu.:1996   3rd Qu.: 9.784
                                      Max.   :2014   Max.   :36.503
   Streams          Downloads        Radio.Plays         Rating
Min.   : 0.000   Min.   : 0.000   Min.   : 0.000   Min.   :0.0000
1st Qu.: 2.351   1st Qu.: 1.040   1st Qu.: 0.600   1st Qu.:0.0000
Median : 4.200   Median : 2.200   Median : 2.570   Median :0.5120
Mean   : 4.629   Mean   : 2.970   Mean   : 3.985   Mean   :0.4819
3rd Qu.: 6.307   3rd Qu.: 4.184   3rd Qu.: 5.884   3rd Qu.:0.5710
Max.   :25.545   Max.   :19.780   Max.   :24.393   Max.   :4.4600
```

Standard Deviation:

```
     Sales        Year     Streams   Downloads Radio.Plays      Rating
 4.5557147  19.7596764   3.2737759   2.6687443   4.3544187   0.6247193
```

Correlation:

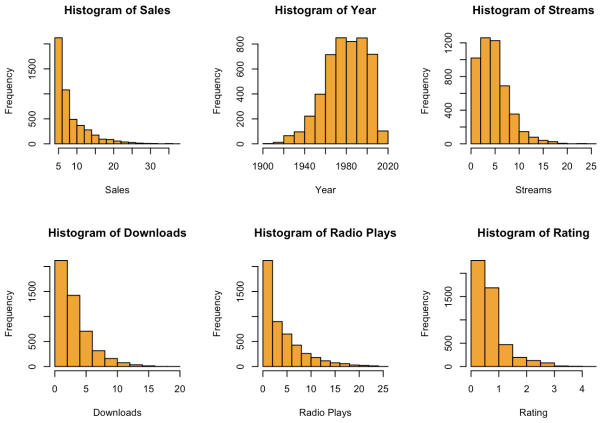
	Sales	Year	Streams	Downloads	Radio.Plays	Rating
Sales	1.0000000	0.06849930	0.55614871	0.66219628	0.5846979	0.38109537
Year	0.0684993	1.00000000	-0.30031068	0.02033752	0.3645831	0.33389305
Streams	0.5561487	-0.30031068	1.00000000	0.25548261	-0.1384287	0.01907206
Downloads	0.6621963	0.02033752	0.25548261	1.00000000	0.4241359	0.02291211
Radio.Plays	0.5846979	0.36458314	-0.13842869	0.42413588	1.0000000	0.27628983
Rating	0.3810954	0.33389305	0.01907206	0.02291211	0.2762898	1.00000000

2.2 Dataset Glossary and Variables

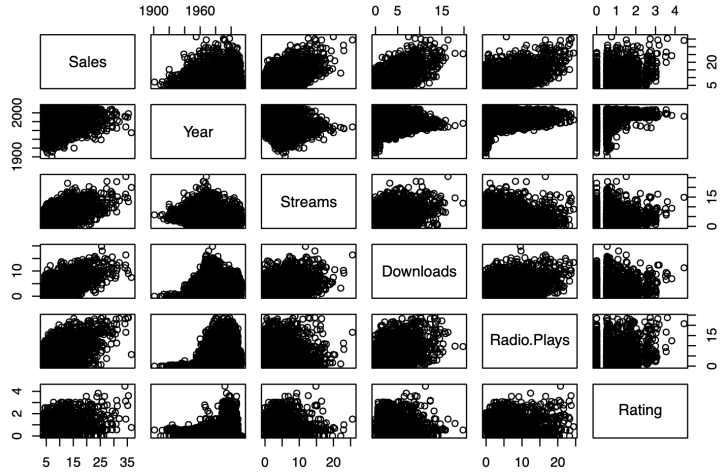
- **Artist:** The name of the song performers.
- **Title:** The title of the song
- **Year:** The year in which the song was released.
- **Sales:** The sales of each song in millions of dollars.
- **Streams:** The total streams of each song have received in millions.
- **Downloads:** The total downloads for each song in millions.
- **Radio Plays:** The number of times when radio plays the song, in millions.
- **Rating:** The numerical rating of the song from 0 to 5.

2.3 Distribution and Relationships Among the Variables

Use histogram to show the distribution of each variables:



Use correlation plot to show the relationship between each variables:



3. Methodology

3.1 Data Filtering

- Since just listing the releasing year might be meaning less, we transformed this column to “How many years has this song been released”, i.e using equation 2024 - year released instead
- We cleared rows where 0s are represented since 0 means the data is absent

3.2 Modeling

First, summary the original model:

From the summary table, we find that Year has a p-value greater than 0.05, which will be considered as an insignificant value, and conduct an alternate model. We decide to use a reduced model without Year.

```
lm1 <- lm(Sales ~ Year + Streams + Downloads + Radio.Plays + Rating, data = numr_data)
mod <- summary(lm1)
mod
```

```
##
## Call:
## lm(formula = Sales ~ Year + Streams + Downloads + Radio.Plays +
##     Rating, data = numr_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.7639 -0.8192 -0.1153  0.5411 12.5493
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.598763   0.138656   4.318 1.64e-05 ***
## Year        -0.005945   0.003103  -1.916   0.0555 .
## Streams      0.773008   0.012454  62.068  < 2e-16 ***
## Downloads    0.594402   0.016242  36.596  < 2e-16 ***
## Radio.Plays  0.510166   0.009544  53.453  < 2e-16 ***
## Rating       1.332881   0.061940  21.519  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.636 on 2108 degrees of freedom
## Multiple R-squared:  0.9139, Adjusted R-squared:  0.9137
## F-statistic:  4474 on 5 and 2108 DF,  p-value: < 2.2e-16
```

Alternate model:

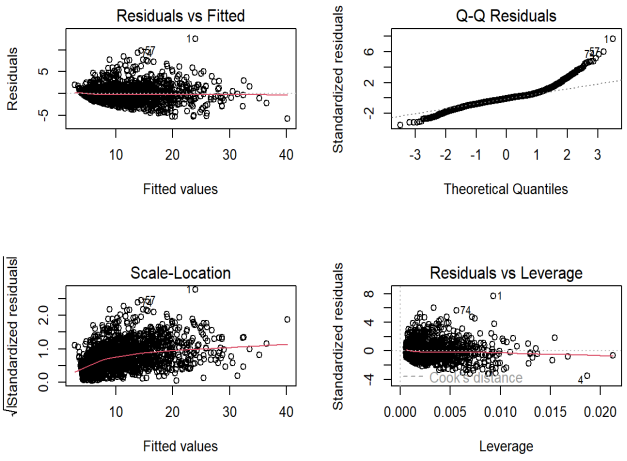
- From the anova table, we noticed that the p-value is greater than 0.05, we do not reject the null hypothesis, thus we suggest an alternate model is a better choice.
- Summary table and diagnostic plot of alternate model:

```
lm2 <- lm(Sales ~ Streams + Downloads + Radio.Plays + Rating, data = numr_data)
anova(lm2, lm1)
```

```
## Analysis of Variance Table
##
## Model 1: Sales ~ Streams + Downloads + Radio.Plays + Rating
## Model 2: Sales ~ Year + Streams + Downloads + Radio.Plays + Rating
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1   2109 5655.0
## 2   2108 5645.2  1    9.8283 3.6701 0.05553 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(lm2)
```

```
##
## Call:
## lm(formula = Sales ~ Streams + Downloads + Radio.Plays + Rating,
##     data = numr_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.7001 -0.8564 -0.1117  0.5741 12.5044
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.390921   0.086397   4.525 6.38e-06 ***
## Streams     0.765179   0.011772  65.000  < 2e-16 ***
## Downloads   0.587274   0.015820  37.121  < 2e-16 ***
## Radio.Plays 0.514835   0.009234  55.756  < 2e-16 ***
## Rating      1.366448   0.059448  22.986  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.637 on 2109 degrees of freedom
## Multiple R-squared:  0.9137, Adjusted R-squared:  0.9136
## F-statistic:  5585 on 4 and 2109 DF,  p-value: < 2.2e-16
```



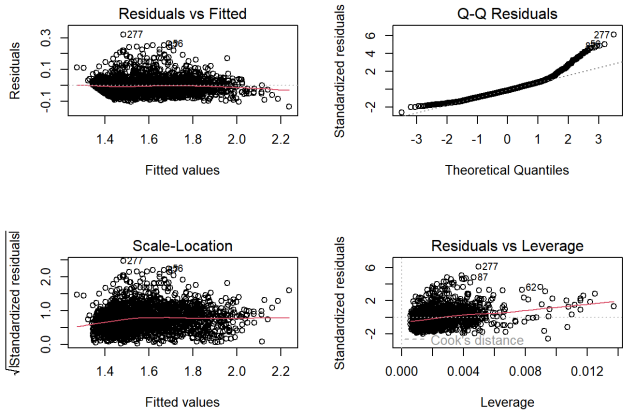
- From the plots, there is some violation to the assumption. There are some outliers that probably affect the regression model. From the Q-Q plot, the head and tail both have deviations from the normality of error. Thus, we consider a transformation.

```
## bcPower Transformations to Multinormality
##    Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## Y1    0.0847          0.08       0.0333       0.1360
## Y2    0.4470          0.45       0.4156       0.4784
## Y3    0.1408          0.14       0.1003       0.1814
## Y4    0.3632          0.33       0.3277       0.3986
## Y5    0.3240          0.33       0.2902       0.3578
##
## Likelihood ratio test that transformation parameters are equal to 0
##  (all log transformations)
##                                     LRT df       pval
## LR test, lambda = (0 0 0 0 0) 1623.185  5 < 2.22e-16
##
## Likelihood ratio test that no transformations are needed
##                                     LRT df       pval
## LR test, lambda = (1 1 1 1 1) 4945.557  5 < 2.22e-16
```

- According to the box-cox, the summary table and diagnostic plot of trans-model is:

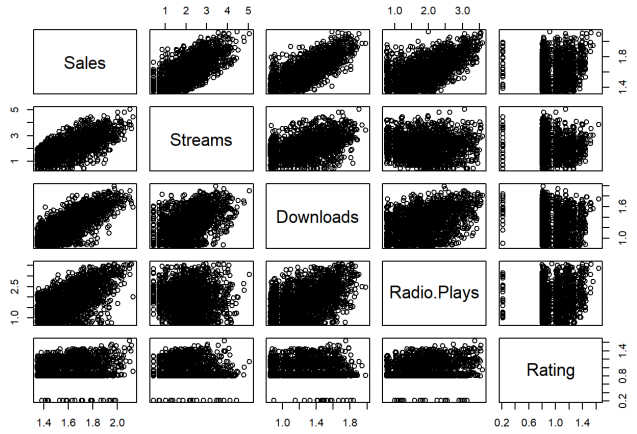
```
summary(transxy1)
```

```
##
## Call:
## lm(formula = new_data$Sales ~ new_data$Streams + new_data$Downloads +
##     new_data$Radio.Plays + new_data$Rating)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.13487 -0.03411 -0.00670  0.02543  0.32038
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          0.782424   0.008059   97.09   <2e-16 ***
## new_data$Streams     0.109193   0.001709   63.90   <2e-16 ***
## new_data$Downloads   0.226078   0.005699   39.67   <2e-16 ***
## new_data$Radio.Plays 0.111443   0.002074   53.74   <2e-16 ***
## new_data$Rating      0.114187   0.006117   18.67   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0527 on 2109 degrees of freedom
## Multiple R-squared:  0.9033, Adjusted R-squared:  0.9032
## F-statistic:  4927 on 4 and 2109 DF,  p-value: < 2.2e-16
```



- The new plots show less violation of assumptions. There are less outliers, and the Q-Q plot looks better even though there is still a heavy tail.
- From the scatter plot on the right, compared with the initial model, we can find the improvement for Streams, Downloads and Radio.Plays, but there are some blank areas of Rating.
- Check the vif at the bottom:
  We performed the model selection even though the VIF score is low.

```
pairs(Sales ~ Streams + Downloads + Radio.Plays + Rating, data = new_data)
```




```
vif(transxy1)
```

```
##     new_data$Streams   new_data$Downloads new_data$Radio.Plays
##             1.356995             1.741793             1.435728
##      new_data$Rating
##             1.079401
```

- Model Selection: By finding All Subsets, the Full Model is the best with highest Radj^2 and lowest AIC, AICc, BIC.

```
##   Subset_size     Radj2       AIC       AICc       BIC
## 1           1 0.5997773  -9442.092  -9442.081  -9430.779
## 2           2 0.8239861 -11177.656 -11177.637 -11160.687
## 3           3 0.8872068 -12117.402 -12117.374 -12094.777
## 4           4 0.9031550 -12438.672 -12438.632 -12410.390
```

```
##               Df Sum of Sq     RSS    AIC
## + Radio.Plays  1   10.2641  6.8249 -12117
## + Rating       1    3.2119 13.8771 -10617
## <none>                     17.0890 -10179
##
## Step:  AIC=-12117.4
## Sales ~ Downloads + Streams + Radio.Plays
##
##          Df Sum of Sq    RSS    AIC
## + Rating  1   0.96777 5.8571 -12439
## <none>                6.8249 -12117
##
## Step:  AIC=-12438.67
## Sales ~ Downloads + Streams + Radio.Plays + Rating
```

- The backward and forward AIC selection all suggest the full model as the best model.

```
## Start:  AIC=-12438.67
## new_data$Sales ~ new_data$Streams + new_data$Downloads + new_data$Radio.Plays +
##     new_data$Rating
##
##                       Df Sum of Sq     RSS    AIC
## <none>                               5.8571 -12439
## - new_data$Rating      1    0.9678  6.8249 -12117
## - new_data$Downloads   1    4.3709 10.2280 -11262
## - new_data$Radio.Plays 1    8.0200 13.8771 -10617
## - new_data$Streams     1   11.3385 17.1956 -10164
```

4.Results and Interpretation

Suggest two or three models (candidates) you have tried.

Transxy1,Transxy2

Determine the "best" predictive model and justify your choice.

To determine the "best" predictive model between Transxy1 and Transxy2, we would typically compare their performance based on several criteria, such as:

Adjusted R-squared: This metric adjusts the R-squared for the number of predictors in the model and the sample size, providing a better measure of fit for models with a different number of predictors. The higher the adjusted R-squared, the better the model explains the variance of the dependent variable while penalizing for adding less informative predictors.

Residual Standard Error (RSE): This measures the average amount that the response will deviate from the true regression line. A lower RSE indicates a better fit.

Significance of coefficients: In both models, all coefficients are highly significant ($p < 2.2e-16$), indicating that each predictor contributes meaningfully to the model.

F-statistic: A higher F-statistic indicates that the model is a better fit for the data. However, the F-statistic should be evaluated in conjunction with the p-value.

Predictive performance: This involves how well the model generalizes to new, unseen data. Techniques such as cross-validation can be helpful in assessing this.

From the summaries provided, here's a comparison:

Transxy1:
Adjusted R-squared: 0.9032
Residual standard error: 0.0527

Transxy2:
Adjusted R-squared: 0.8957
Residual standard error: 0.08098

Based on these metrics, Transxy1 has a higher adjusted R-squared value and a lower residual standard error, suggesting that it is the better model in terms of both fit and predictive accuracy.

The F-statistic is very high for both models, indicating strong predictive power, but since Transxy1 has a higher adjusted R-squared and a lower RSE, it is the preferred model according to these results. The model selection using both stepwise and finding all subset approaches all suggest the full model as the model of best fit, so keep Transxy1 as the best choice.

5 Discussion

This project explores the impact of various factors on music sales in the digital era by utilizing the "Top Songs of the World" dataset from Kaggle. The study employs multiple regression analysis to determine the influence of streams, downloads, radio plays, and ratings on music sales. With comprehensive data including sales figures, streaming counts, and other relevant metrics, the research aims to reveal the dynamics that drive music popularity and success in today's digital landscape. The report is structured to introduce the study, detail the methodology, present findings, and discuss implications, ultimately providing stakeholders with key insights for understanding and navigating the complexities of the digital music market.

The methodology includes data filtering to account for the relevance of a song's release year by calculating the number of years since its release and removing rows with absent data (marked by zeros). An initial regression model indicated the insignificance of the "Year" variable, prompting the adoption of a reduced model excluding it. Diagnostic tools and plots revealed some outliers and deviations from error normality, leading to a transformation of the model. The transformed model showed improved assumption compliance, with less influence from outliers and a better fit indicated by Q-Q plots. The project also discusses the alternate model's suitability, and improvement in model diagnostics, and concludes with a validation of the model's variables through the variance inflation factor (VIF) check, confirming that no further model selection was necessary.

We didn't choose the model, but the project does reflect real-world scenarios, particularly how digital consumption has become a cornerstone of the music industry. It mirrors the trends noted by market analysts and industry observers, highlighting the shift toward streaming as a primary source of revenue for artists and record labels. This trend has only intensified with the pandemic limiting live event revenues, making streaming a vital part of the industry's financial backbone.

Deloitte's insights into the music industry affirm the significance of streaming, especially during the COVID-19 pandemic, which has seen the cancellation of live events and a consequent reliance on digital platforms for revenue. The reports by Deloitte detail the accelerated growth in music revenues fueled by streaming from 2015 to 2019, which aligns with the model's indications of streaming's positive impact on sales.

While the analysis provides substantial insights, it is not without limitations. The model does not account for all possible variables that may influence music sales, such as cultural shifts, individual artist influence, changes in consumer purchasing behavior, and the impact of marketing campaigns.

To refine the predictive accuracy, future iterations of the project could incorporate additional variables, including social media engagement metrics, global economic indicators, and more granular consumer data. The use of more sophisticated machine learning models that can adjust to non-linear relationships and interactions between variables could also enhance the robustness of the findings.

The project's strength lies in its quantitative approach and the high level of explanatory power of the chosen model. It successfully identifies and quantifies the relationship between key factors and sales, which can inform strategic decision-making in the music industry. The primary weakness is the static nature of the model, which may not adapt well to rapid changes in the music industry.

Reference:

“How Streaming Is Changing the Music Industry.” Deloitte Insights, www2.deloitte.com/us/en/insights/industry/technology/how-streaming-is-changing-the-music-industry.html. Accessed 7 Mar. 2024.