Final Project - Step 2 (20 Points)

PSTAT126: Regression Analysis

Ali Abuzaid

STUDENT NAME

- AARTI GARAYE (aartigaraye)
- MOIRA KEATING (mskeating)
- YIWEN XIAO (yiwenxiao)
- SYLVIA LI (sylvia_li)
- SIDDHARTH SINGH (siddharthsingh)

Due Date

The deadline for this step is **November 8, 2024**.

Instructions

The goal of this step is to develop clear research questions and hypotheses based on your selected dataset and to conduct a thorough Exploratory Data Analysis (EDA). This process will set the foundation for your later analysis and insights.

1 Step 2: Research Questions, Hypotheses, and Exploratory Data Analysis (EDA)

1.1 Research Objective

How can artists use our model to strategically structure their music and plan the release dates to maximize their structure?

1.2 Research Questions

Question 1

How does the number of playlists a song is in influence the number of streams?

Question 2

Out of acousticness %, number of playlists, bpm, spotify charts, which variables are statistically significant for our model to predict the number of streams?

Question 3

How does release month impact the number of streams?

1.3 Hypotheses

Hypothesis 1

Null Hypothesis: H_0: $\beta_1 = 0$ The number of playlists a song is in (predictor variable) and the number of streams (response variable) have no linear relationship.

Alternate Hypothesis: H_A: $\beta_1 \neq 0$ The number of playlists a song is in (predictor variable) and the number of streams (response variable) have some linear relationship.

Hypothesis 2

Null Hypothesis: H_0: $\beta_1 = 0, \beta_2 = 0, ..., \beta_p = 0$ None of the variables listed have a statistical impact on the number of streams.

Alternate Hypothesis: H_A: $\beta_1 \neq 0, \beta_2 \neq 0, ..., \beta_p \neq 0$ At least one of the variables listed above have a statistical impact on the number of streams.

Hypothesis 3

Null Hypothesis: H_0: $\beta_1 = 0, \beta_2 = 0, ..., \beta_p = 0$ None of the months are statistically significant to impact the number of streams.

Alternate Hypothesis: H_A: $\beta_1 \neq 0, \beta_2 \neq 0, ..., \beta_p \neq 0$ At least one of the months are statistically significant to impact the number of streams.

1.4 Exploratory Data Analysis (EDA)

1.4.1 Data Cleaning

The variables we chose were the number of playlists, charts, bpm, and acousticness percentage. The reason we chose to come up with hypothesis tests for these variables is because we thought that they would have the most impact on the number of streams. We thought that these variables would have the strongest correlation, which would make it easier for us to make linear regression models. The graphs below shocked us.

We cleaned the data set with the help of R and excel. First using R, we omitted all of the rows that had at least one NA or empty value. We then used the set.seed function to randomly select 500 tracks so that our data would not be skewed. Then, we exported the dataset into Excel and deleted all of the columns that we unnecessary and redundant. Once we did that, we converted the track names to index variables because not all of them were in English, which gave us a lot of discrepancies. This also helps to reduce biases by looking at artists track and name.

In our experiment, we will not be removing outliers because we are unaware of the circumstances making it unjustifiable to just omit them. Doing so would change the summary statistics significantly, which can potentially affect the central measures of tendency and the spread of the data. This would make it harder to come up with the linear regression models because our variables might be less correlated with each other without the extreme values.

1.4.2 Descriptive Statistics

```
library(readxl)
library(knitr)
library(kableExtra)
library(dplyr)
spotify_data <- read_excel("/Users/aarti/Downloads/new_Spotify_data.xls")
spotify_data$streams <- as.integer(spotify_data$streams)
spotify_data$released_month <- as.character(spotify_data$released_month)
summary(spotify_data)
```

track_index_number	artist_count	released_year	released_month	released_day	in_spotify_playlists	in_spotify_charts	streams
Length:500	Min. :1.000	Min. :1942	Length:500	Min. : 1.00	Min. : 31.0	Min. : 0.00	Min. :2.762e+03
Class :character	1st Qu.:1.000	1st Qu.:2020	Class :character	1st Qu.: 5.00	1st Qu.: 893.5	1st Qu.: 0.00	1st Qu.:1.370e+08
Mode :character	Median :1.000	Median :2022	Mode :character	Median :13.00	Median : 2180.5	Median : 2.00	Median :2.565e+08
NA	Mean :1.536	Mean :2018	NA	Mean :13.71	Mean : 5018.1	Mean : 11.31	Mean :4.369e+08
NA	3rd Qu.:2.000	3rd Qu.:2022	NA	3rd Qu.:21.25	3rd Qu.: 5215.0	3rd Qu.: 16.00	3rd Qu.:5.962e+08
NA	Max. :8.000	Max. :2023	NA	Max. :31.00	Max. :51979.0	Max. :147.00	Max. :2.123e+09

NA	NA	NA	NA	NA	NA	NA	NA's :11	

bpm	key	mode	danceability_%	valence_%	energy_%	acousticness_%	instrumentalness_%	liveness_%
Min. : 65.00	Length:500	Length:500	Min. :23.00	Min. : 4.00	Min. :14.00	Min. : 0.00	Min. : 0.000	Min. : 3.00
1st Qu.: 98.75	Class :character	Class :character	1st Qu.:57.00	1st Qu.:32.75	1st Qu.:53.75	1st Qu.: 6.00	1st Qu.: 0.000	1st Qu.: 9.00
Median :120.00	Mode :character	Mode :character	Median :70.00	Median :52.00	Median :65.00	Median :19.00	Median : 0.000	Median :12.00
Mean :122.40	NA	NA	Mean :67.46	Mean :51.66	Mean :64.04	Mean :27.55	Mean : 2.128	Mean :17.89
3rd Qu.:140.00	NA	NA	3rd Qu.:79.00	3rd Qu.:72.00	3rd Qu.:76.00	3rd Qu.:43.25	3rd Qu.: 0.000	3rd Qu.:23.00
Max. :204.00	NA	NA	Max. :95.00	Max. :97.00	Max. :97.00	Max. :93.00	Max. :91.000	Max. :92.00

By looking at the summary of the key variables, we can see that bpm ranges from 65 to 204 with the median being at around 120 and mean being at 122. The in_spotify_playlist variable is how many Spotify playlists the track is in. These values range from 31 to 51,979, with the median being at 2,180, but the mean being at 5,018, which means that there are a lot of extremes closer to the maximum. Streams range from 2,762 to 2,123,000,000 with the median being at 256,500,000 and mean being at 436,900,000. The danceability % variable tells us the suitability of the song for dancing, and it ranges from 23% to 95%, with the median being at 70% and mean being at 67.47%. The valence % tell us the positivity of the song's musical content, and this ranges from 4% to 97% with the median being at 52% and mean being at 51.66%. Energy % tells us the perceived energy level of the song, which ranges from 14% to 97% with the median being 65% and mean being 64%. The acousticness % measures the acoustic sound presence in the song, and in this case, acousticness % 0% to 93%, with the median being 19% and mean being 27.55%, telling us that this variable's data is slightly skewed to the right.

```
library(ggplot2)

ggplot(data = spotify_data, mapping = aes(x=factor(released_month, levels=as.character(1:12)), y=streams)) +
  geom_boxplot() +
  theme_minimal() +
  xlab("Months") +
  ggtitle("Boxplot of streams for every month")
```

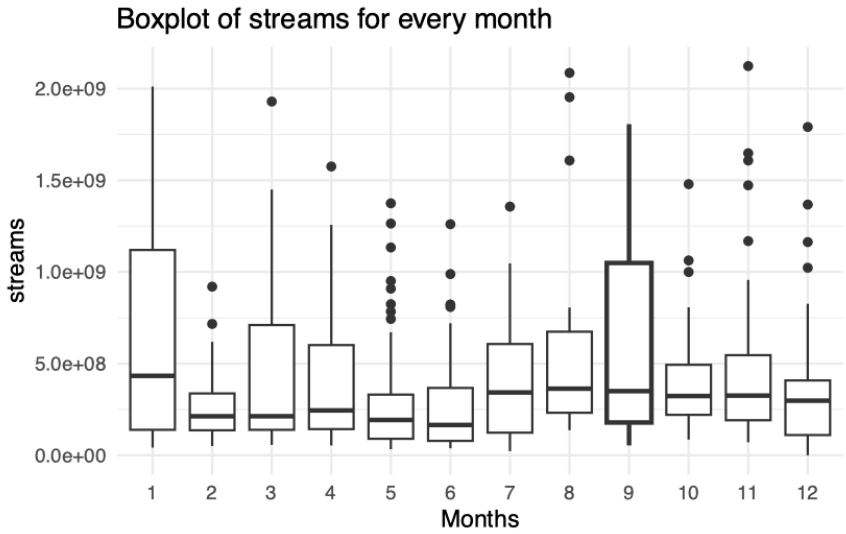

Figure 1: Side by side boxplot of streams every month. We can see some outliers in them. January has the highest median whereas June has the lowest median number of streams.

1.4.3 Data Visualization

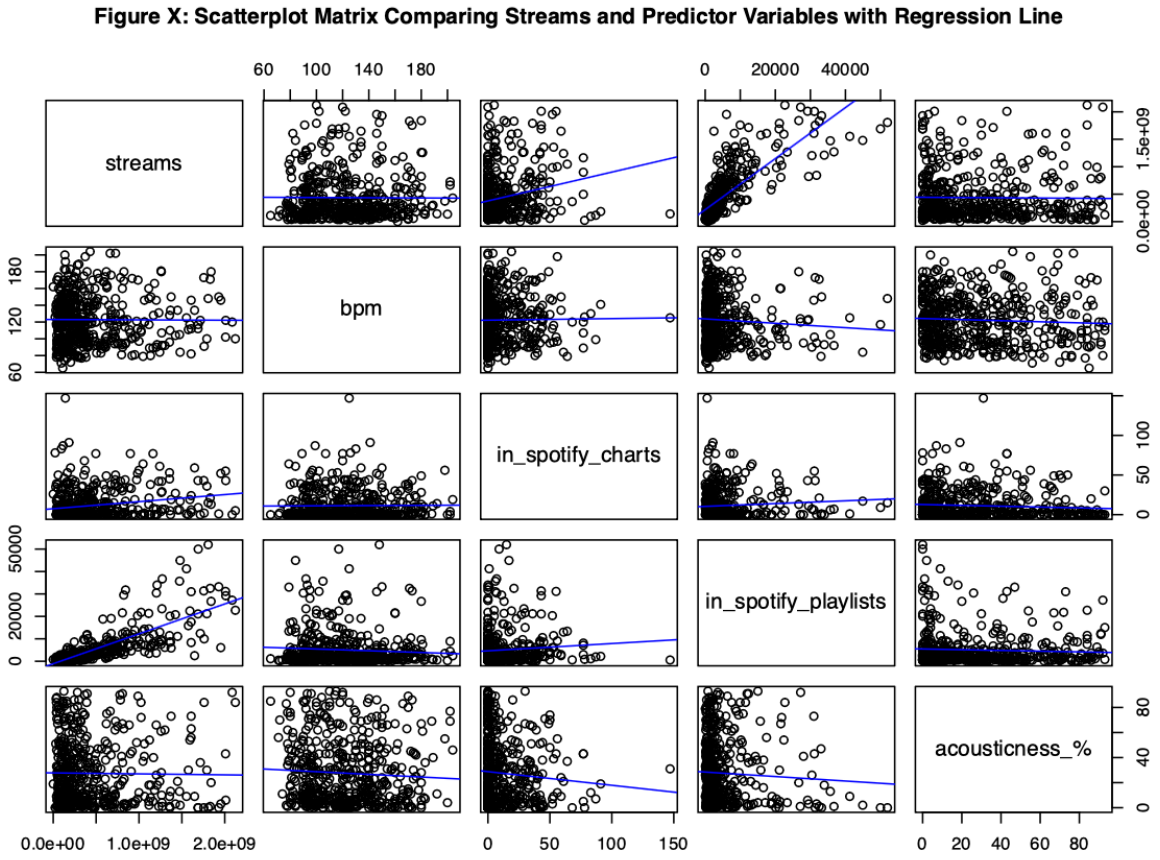

Figure 2: Scatterplot matrix showing relationships between streams, BPM, artist count, acousticness percentage, Spotify charts, and Spotify playlists. Insights: On average, **streams** increase as being **in_spotify_playlists** increases but there is variation. The relationships between all pairs of predictors appear to be very weak.

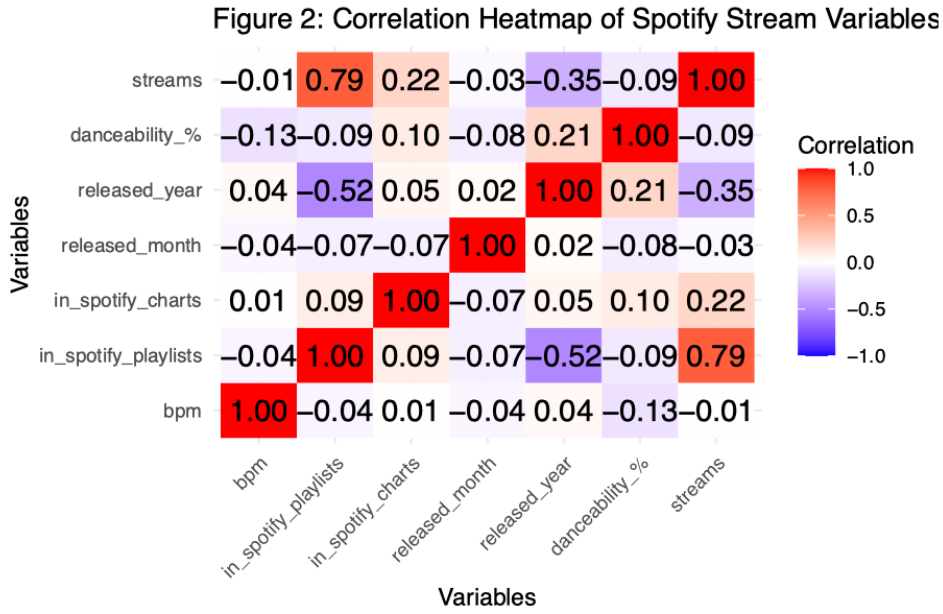

Figure 3: The positive relationship between **streams** and **in_spotify_playlists** suggests that being in more playlists and high streams are closely related, with the most notable correlation. **Acousticness_%** and **bpm** show very weak correlations to streams and **in_spotify_charts** shows a positive correlation but it is still quite weak. There appears to be a somewhat negative correlation between **released_year** and number of streams but it's still relatively weak.

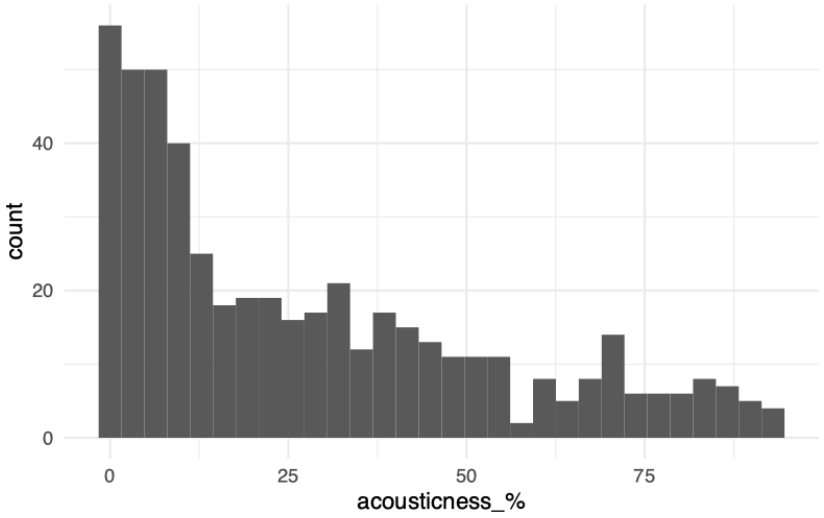

Figure 4: A histogram of acousticness % reveals a distribution that is not normal and skewed right. This means that majority of the songs have 0% acousticness.

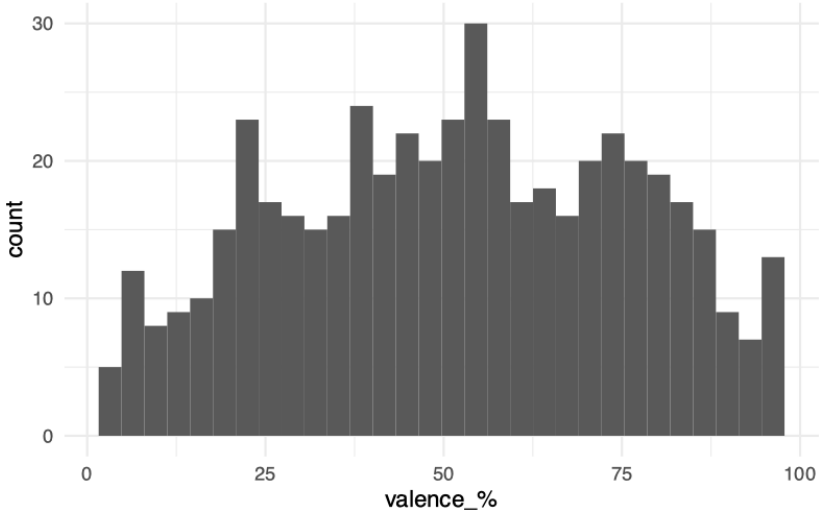

Figure 5: A histogram of valence % reveals a roughly normal distribution with the mean centered around 50% and the tails towards 0 and 100%.

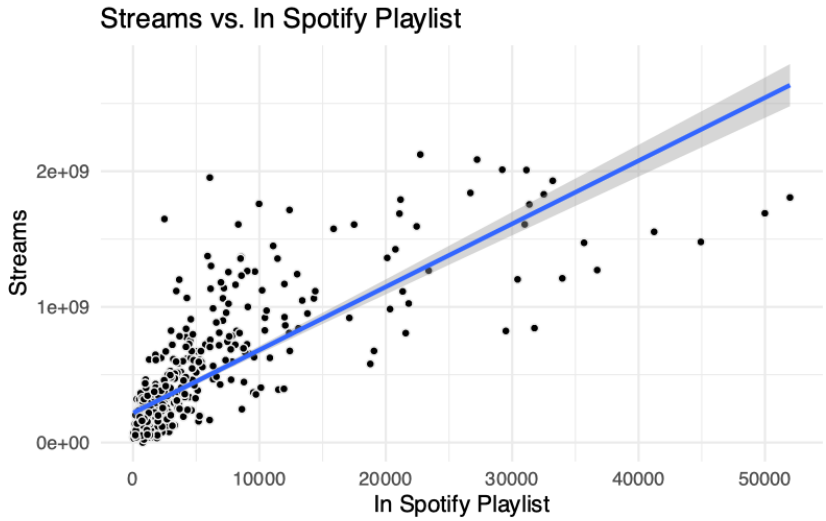

Figure 6: The scatter plot shows a moderately strong positive relationship, as In Spotify Playlist increases, Streams increases. The variance increases as In Spotify Playlist increases, and there are a few identifiable outliers.

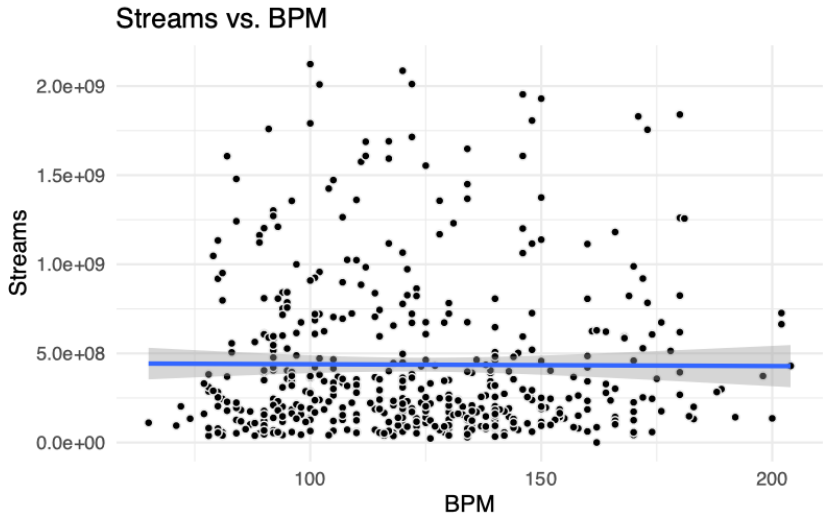

Figure 7: We expected a positive linear relatinship, but there doesn't seem to be a strong linear relationship between BPM and Streams, and no visible pattern is observed.The variance is large, making it difficult to identify outliers.

1.4.4 Feature Relationships

From the above graphs of different scatterplots and heatmaps, we can see that the strongest positive correlation is between number of playlists a song is in and the number of streams it has. So our initial analysis tells us the the number of playlists a song is in has some statistical significance in determining the number of streams. The heatmap also suggests a negative correlation between the release year and streams.

1.4.5 Distribution Comparisons

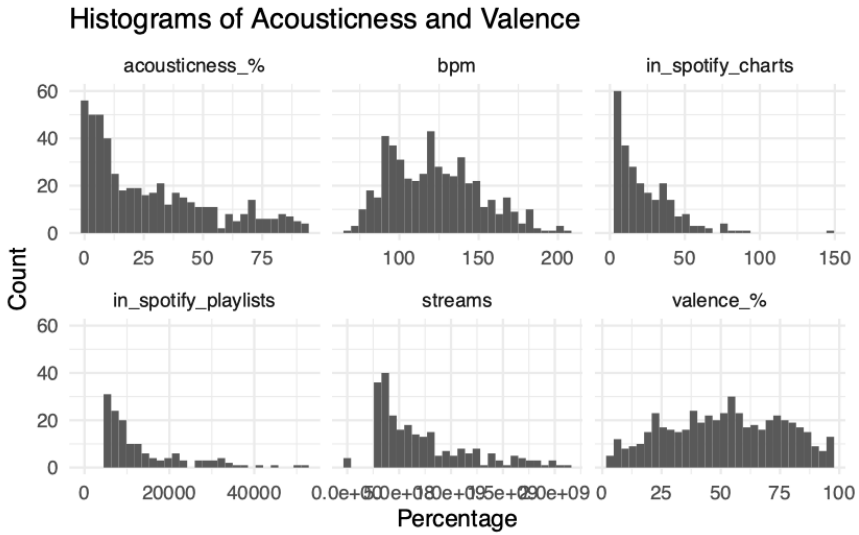

Figure 8: It appears that acousticness, playlists, charts, and streams exhibit a right-skewed distribution, while BPM and valence follow a more normal distribution.

1.4.6 Skewness and Kurtosis

Table 3: Skewness and Kurtosis in different variables

	Skewness	Kurtosis
bpm	0.4447840	-0.4478829
in_spotify_playlists	2.9576039	9.9085995
in_spotify_charts	2.4460095	8.7787017
acounsticness_%	0.8743688	-0.3336743

For BPM, the distribution has a skewedness of 0.445 and a kurtosis of -0.448. These values reflect a mostly symmetrical distribution that is close to normal in terms of being not too flat or peaked. More normal distributions have skewedness and kurtosis values close to 0. In contrast, the In_Spotify_Playlists variable has a skewedness of 2.96 and a kurtosis of 9.91. These values fall well outside the commonly accepted values for normal distributions, -2 to 2, reflecting that the

distribution for number of Spotify playlists a song is in is not normally distributed. This also applies for the variable In_Spotify_Charts, with a skewedness of 2.45 and a kurtosis of 8.78. The distribution for acousticness % is closer to normal, with a skewedness of 0.874 and a kurtosis of -0.334.

1.4.7 Documentation

The dataset is interesting and relevant for analysis because it includes variables with both skewed and normal distributions. In addition, some of those variables exhibit linear correlations, which can provide deeper insights into the relationships between number of streams on Spotify and those musical characteristics. This model will help musicians to know which aspects to focus on and how to best release their work to optimize the streams.