A Predictive Model for Spotify Artist Growth Anchored in Christmas

Humphrey Boadi, Josh Bates, Shane Schwartz

Abstract

In this project, we went through the process of creating a time series analysis and created predictive models of artists who have massively popular Christmas music and for artist(s) who do not with the Monthly Listeners data sources from Spotify. The aim of our project is to understand and analyze the ties between popular holiday music and artists and the time of year, focusing on Christmas. We hypothesized that the time of year does have a relationship with the amount of monthly listeners certain artists get. This turned out to be true for certain artists with popular Christmas music and not for otherwise.

1.0 Introduction:

In the contemporary music industry, the marriage of marketing and analysis stands as the paramount factor influencing success. The strategic use of mathematics and statistical modeling plays a pivotal role, holding the key to orchestrating the timing of significant events and accurately predicting trends, the ramifications of which extend beyond mere monetary quantification. A prime example of this synergy is the ability to forecast an artist's surge in popularity, whether through the nuanced SARIMA modeling or the straightforward elegance of linear regression analysis. In essence, the expanding repository of data in the musical landscape opens up avenues for deploying a spectrum of effective methods, providing valuable insights into the inherently dynamic and financially consequential music industry In step with the contemporary technological landscape, machine learning and regression analysis emerge as indispensable tools, offering a plethora of possibilities for predicting trends and comprehending the profound impact that individual artists and music genres can exert. The swift accessibility and user-friendly nature of numerical data and statistics underscore the transformative potential of these analytical techniques. Much like the meticulous collection of insights into various fields, including the study of diseases, the music industry has amassed a wealth of information concerning artists on platforms like Spotify. This encompasses everything from understanding the mechanics of popularity to discerning the influence of critical events. In this pursuit, our aim is to systematically gather pertinent information, leveraging the power of statistical analysis to generate well-founded predictions about creators—an invaluable reservoir of knowledge poised to illuminate the path through the ever-evolving landscape of the music industry.

1.1 Data:

The dataset under scrutiny was meticulously obtained from SpotifyStats.com, an online platform distinguished by its exclusive access to some otherwise restricted data from the Spotify API. Regularly updated on a daily basis, this comprehensive dataset systematically gathers monthly listener information directly from the Spotify platform. A notable feature of this dataset is its precision, providing an exact account of monthly listener totals for each artist. This historical depth extends back to 2018, with select instances showcasing data from 2020. This chronological breadth not only ensures a thorough representation of the music landscape over time but also sets the stage for precise modeling. The daily updates and direct extraction from Spotify contribute to the reliability and timeliness of the dataset however, as of December 7th, the website creator has taken down the website due to a heavy increase in people opening the website as a result of Spotify Wrapped being released causing the site to repeatedly crash. When opening the website users will be met with the message "Thanks for visiting SpotifyStats.Com

I've taken a decision to shut down the site to new users for the time being. If you want to see the basic Spotify listings for your most listened to songs, then login below.

If you have previously registerd on the site and are just looking to re-login, then again the Login button below should work.

Head to our Facebook page for the latest news.

Thanks! Dave"

By encompassing monthly listener metrics for every artist, the dataset was a valuable resource for accurate modeling endeavors. Its extensive time span, dating back to 2018, allows researchers to discern trends, patterns, and shifts in listener engagement over the years. In instances where data is available only from 2020, it still provides a contemporary snapshot, enabling analyses that capture recent developments.

1.2 Approach:

It is widely acknowledged, perhaps subconsciously, that certain artists experience significant surges in popularity during specific times of the year, such as Christmas or the holiday season in general. Recognizing that different artists are uniquely affected and exhibit distinct growth patterns; we embarked on a fascinating exploration using SARIMA and linear regression models. These models were designed to unveil the strength and repeatability of growth trends exhibited by various creators. Certain alternatives such as finding a Pearson or Spearman correlation coefficient would not have worked due to the nature of our data having a strong repeating pattern. The SARIMA model, with its capacity for providing accurate information on trend strength and repeatability over time, was instrumental in uncovering nuanced patterns. Additionally, depending on the artist's characteristics, linear regression was employed to make single-point numerical predictions, offering a contrast to the overall strength of patterns. The resulting values were skillfully plotted in conjunction with time and average listeners to scrutinize potential relationships with external factors such as the time of year, artist type, and the corresponding level of popularity.

Our focus zeroed in on month-by-month peak monthly listeners, as we believed this approach would best showcase patterns and results derived from our SARIMA and linear regression models, offering a comprehensive display of trend associations.

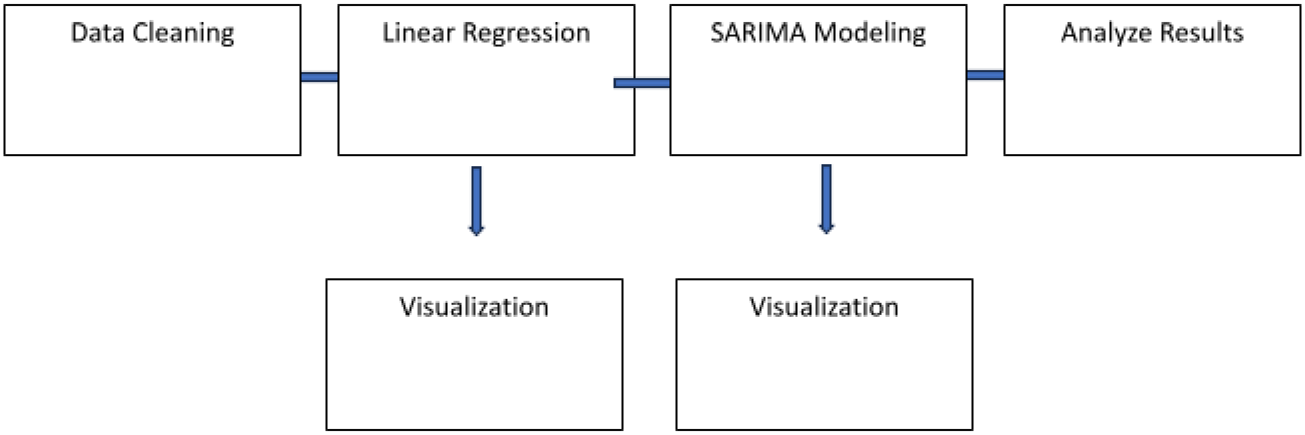
1.3 Summary and Insights:

Our in-depth analysis, fueled by advanced mathematical modeling, unveils key insights into the dynamics of artist popularity and its connection to external factors. Leveraging SARIMA and linear regression models, we've identified nuanced patterns in the growth trajectories of artists, especially during seasonal peaks like Christmas. SARIMA's prowess in capturing trend strength over time highlights the dynamic nature of artist popularity. Linear regression adds a layer, allowing us to make precise predictions and contrast overall pattern strength across diverse artist profiles. Focusing on month-by-month peak listenership, our findings strongly indicate a significant association between artist types and external events, such as the time of year. This sheds light on how certain genres or profiles experience more pronounced popularity fluctuations during specific seasons. In essence, our research provides a meticulous roadmap for industry stakeholders, emphasizing the importance of data-driven strategies to navigate the ever-evolving music landscape with a keen eye on the nuanced factors shaping artist popularity.

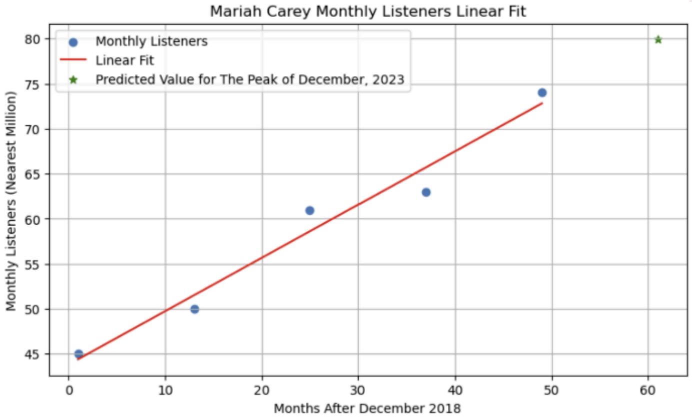
2.0 Methods

The raw data utilized in this study was manually scraped from the previously mentioned website, spotifystats.com resulting in eight CSV files, each containing 2-3 features. Half of these files encompass Peak Monthly listeners for artists month by month, specifying the corresponding months. The remaining four files mirror these features but include an additional one, indicating the month by an index. This indexing facilitated streamlined graphing for linear regression analysis. All monthly listener data considered in our analysis spans from December 2018 to the present time, with a few exceptions for certain artists. While the original data provided daily listener information, we focused specifically on the month-by-month maximum totals for our analytical purposes. This refined dataset serves as the foundation for our mathematical modeling and statistical analysis, ensuring a comprehensive and insightful exploration of artist popularity trends.

The workflow is pictured below on the top of figure A and will be explained more deeply.

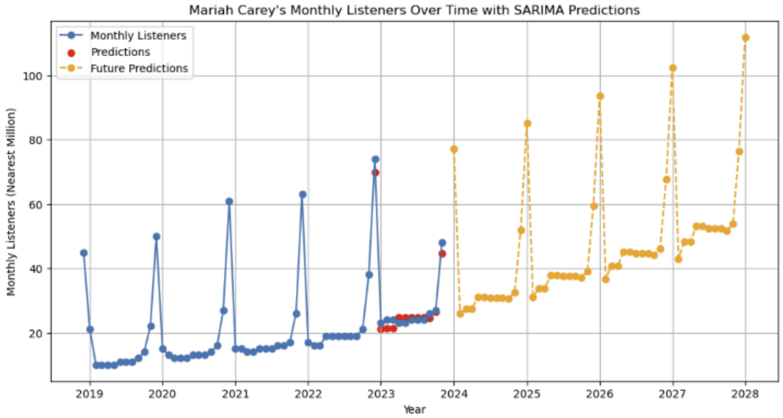


Visualization of Workflow



```
Predicted Monthly Listeners for Peak of December 2023 (Millions): 79.90
```

```
Mean Squared Error: 4.638621396898887
```



Linear Regression of Peak December Listeners for Mariah Carey

SARIMA Prediction Model for Mariah Carey

Figure A. Workflow Diagram and Example Results

2.1 Data Cleaning:

Once we scraped the data manually from our source, we decided to omit all data points besides the peak listeners of each month, as we felt it better isolated and displayed the trends we were looking for. We then read the CSV file using the available functions from the pandas library into a data frame structure and then manipulated it to display the information we desired

Our results for the data from the data cleaning using Mariah Carey as an example can be seen on the bottom of Figure A.

2.2 Linear Fit:

We utilized one method for fitting, and it was linear. Because a big goal of our project was to show a direct, linear correlation, linear fit was the best choice when displaying the data as is. We mapped the monthly listeners on the y axis and the time on the x axis.

$$Y = mx+b$$

Where Y represents the predicted values of the response variable, m represents the slope of the linear function, x represents values for the explanatory variable, and b represents the Y-intercept

2.3 Linear Regression:

Seeing that our initial thought was that there was a big correlation between monthly listeners and time of year we hypothesized that linear regression would provide a very valuable visual due to the fact that we are trying to show single point trends in order to make predictions about the expected amounts of monthly listeners at the peak of each month. The linear regression line would also aid in determining the strength of the correlation in the data between the two axis.

$$\hat{Y} = bX + a$$

Where $\hat{Y}$ represents the predicted values of the response variable, b represents the slope of the linear function, X represents values for the explanatory variable, and a represents the Y-intercept

2.4 SARIMA Modeling:

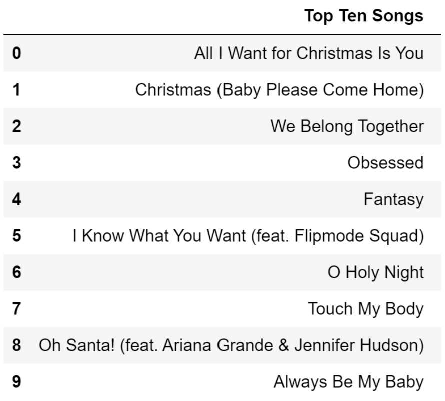
Given our initial hypothesis positing a substantial correlation between monthly listeners and the time of year, we envisioned SARIMA modeling as a potent analytical tool. SARIMA (Seasonal Autoregressive Integrated Moving Average) provides a robust framework for time series forecasting, aligning seamlessly with our objective of revealing single-point trends to predict expected monthly listener counts at peak times.

$$\hat{Y} = bX + a$$

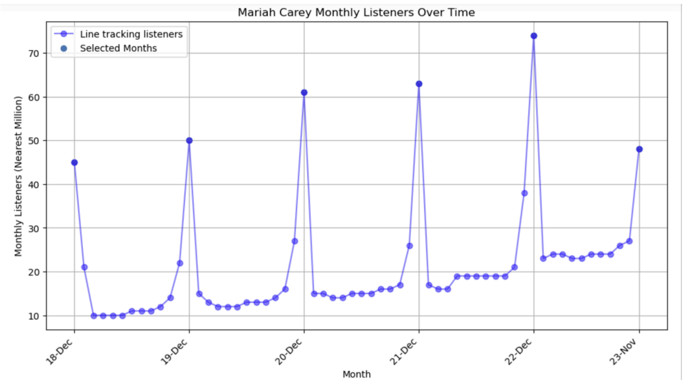
Here, $\hat{Y}$ signifies the predicted values of the response variable, b represents the slope of the linear function, X denotes values for the explanatory variable, and a signifies the Y-intercept. This formula encapsulates the essence of our SARIMA approach where we aim to generate accurate predictions and gauge the strength of the correlation between the temporal axis and monthly listener counts. By leveraging SARIMA, we anticipate illuminating intricate patterns and trends in artist popularity over time, offering valuable insights for both visual representation and predictive modeling.

3.1 Results/Discussion of Christmas Giants

We analyzed three different Christmas Giants, including Micheal Buble, Wham!, and Mariah Carey, but the data for our presentation needed to include Micheal Buble for the sake of time constraints.



	Top Ten Songs
0	All I Want for Christmas Is You
1	Christmas (Baby Please Come Home)
2	We Belong Together
3	Obsessed
4	Fantasy
5	I Know What You Want (feat. Flipmode Squad)
6	O Holy Night
7	Touch My Body
8	Oh Santa! (feat. Ariana Grande & Jennifer Hudson)
9	Always Be My Baby



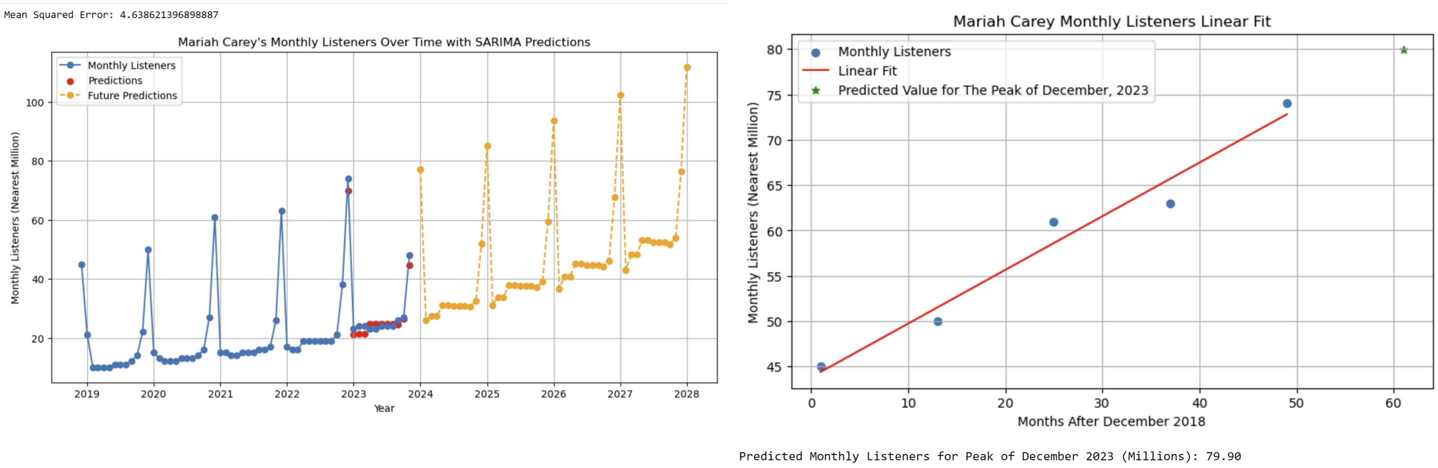
Mariah Carey's Top 10 Songs as of December 4, 2023

Mariah Carey Dot Plot of Monthly Listeners from 2020-2023

Figure B. Mariah Carey Top Songs and Monthly Listeners

The first “Christmas Giant” we analyzed was singer and songwriter Mariah Carey. As mentioned previously, we used the Spotify API (As of December 4th, 2023) to see the artist's Top 10 most popular songs. As seen on the left of Figure B, 4 out of 10 most popular songs are Christmas songs. Mariah Carey’s most popular Christmas work is an album called “Merry Christmas” which came out in 1994, and features smash hits like “All I Want for Christmas is You”. We also created a timeline from 2018-now for all the artists to see their most current public outings including new album releases and tours and such. Carey went on tour in 2018, 2019, 2022 and 2023, with the last two being Christmas tours. She also releases Christmas specials in 2020 and 2021.

The dot plot for Mariah Carey (right of Figure B) shows a relatively low monthly listener count from the months of January through October of the respective years. Then, you see an increase in November and a large spike in the months of December. Then it goes back down and the pattern continues.



```
Mean Squared Error: 4.638621396898887
```

```
Predicted Monthly Listeners for Peak of December 2023 (Millions): 79.90
```

Mariah Carey SARIMA Predictive Model ***Mariah Carey Linear Regression Prediction Model***

Figure C: Predictive Models of Mariah Carey

With our SARIMA Predictive model for Mariah Carey (left of Figure C), we predict the pattern of relatively low monthly listeners in the non-seasonal months and high monthly listeners in the month of December. We used the data from 2021-2023 to train the model. And get a mean squared error. We are led to believe that our SARIMA Predictive Model is accurate because of our low Mean Squared Error of 4.64, which reinforces the accuracy of our predictive model.

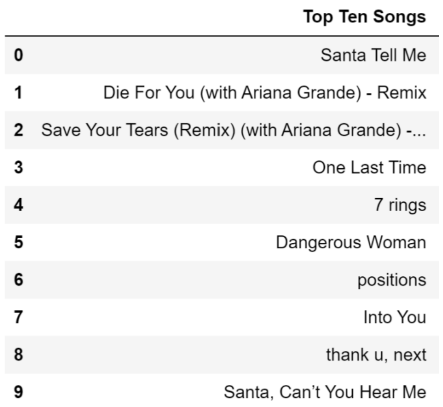
Our linear regression model (right of Figure C) predicts the monthly listeners of Mariah Carey to peak for the year 2023 at 79.9 million monthly listeners

We continued this process of analyzing the patterns of the monthly listeners and creating predictive models for two other Christmas gains, including Micheal Buble and Wham! The data visualizations of the respective artists can be found in the Appendix section. Looking at Figure 1 and the appendix for the grouped bar charts, there is a very significant and consistent trend seen in Mariah Carey's monthly listener count in the seasonal season, especially in December, the month of Christmas (This trend continues with the other Christmas Giants of Micheal Buble and Wham!). Our predictive models showing that these trends will continue to occur for the foreseeable future is also significant in formulating an effective marketing plan to promote the Christmas music of these types of artists.

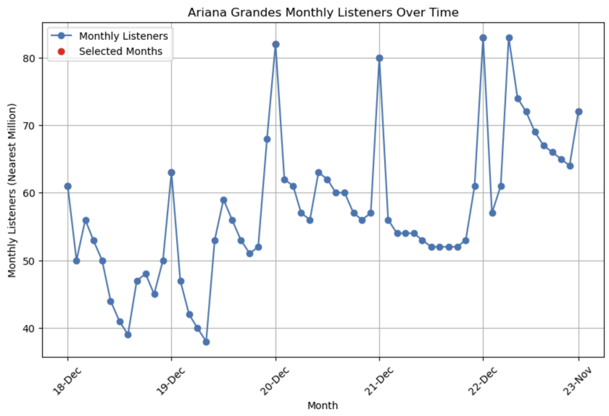
The process of analyzing the patterns shown above and predictive models is part of our time series analysis. The reason for repeating this pattern three times for three different artists is to reiterate the consistency of the pattern of artists like this. For artists that are consistently making appearances currently, whether it be music dropping or tours (like Micheal Buble or Mariah Carey) or artists who are not currently performing but have their music being promoted and advertised (like Wham!), there is still a large spike in the monthly listeners of these artists.

3.2 Results/Discussion of Overall Popular Artists

For the artists we grouped as “Overall Popular Artists”, we analyzed pop sensation Ariana Grande as well as one of the most popular artists of any genre, Drake.

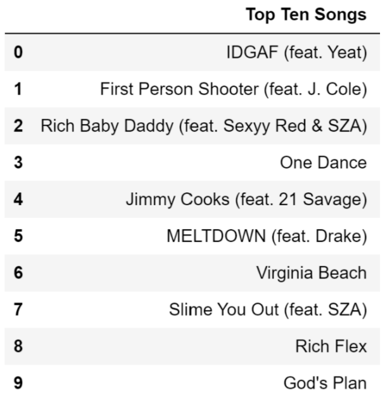


	Top Ten Songs
0	Santa Tell Me
1	Die For You (with Ariana Grande) - Remix
2	Save Your Tears (Remix) (with Ariana Grande) -...
3	One Last Time
4	7 rings
5	Dangerous Woman
6	positions
7	Into You
8	thank u, next
9	Santa, Can't You Hear Me

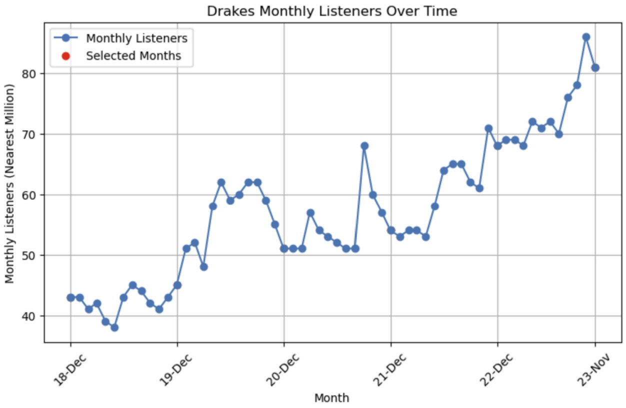


List of Ariana Grande’s Top Ten Songs on Spotify

Dot Plot of Ariana Grande’s Monthly Listeners from 2018-2023



	Top Ten Songs
0	IDGAF (feat. Yeat)
1	First Person Shooter (feat. J. Cole)
2	Rich Baby Daddy (feat. Sexyy Red & SZA)
3	One Dance
4	Jimmy Cooks (feat. 21 Savage)
5	MELTDOWN (feat. Drake)
6	Virginia Beach
7	Slime You Out (feat. SZA)
8	Rich Flex
9	God's Plan



List of Drake’s Top Ten Songs on Spotify

Dot Plot of Drake’s Monthly Listeners 2018-2023

Figure D. Top 10 Songs of Ariana Grade and Drake and Their Monthly Listeners

In Figure D on the top left, Ariana Grande’s top 10 songs list, we can see that 2 out of her Top 10 songs are Christmas songs(“Santa Tell Me” and “Santa, Can’t You Hear Me”). Grande’s most popular Christmas work is a single called “Santa Tell Me” which came out on November 24, 2014. Grande’s timeline consists of releasing albums in 2018, 2019, and 2020, going on tour in 2019, and releasing a Christmas single featuring Kelly Clarkson called “Santa Can’t You Hear Me” in 2021.

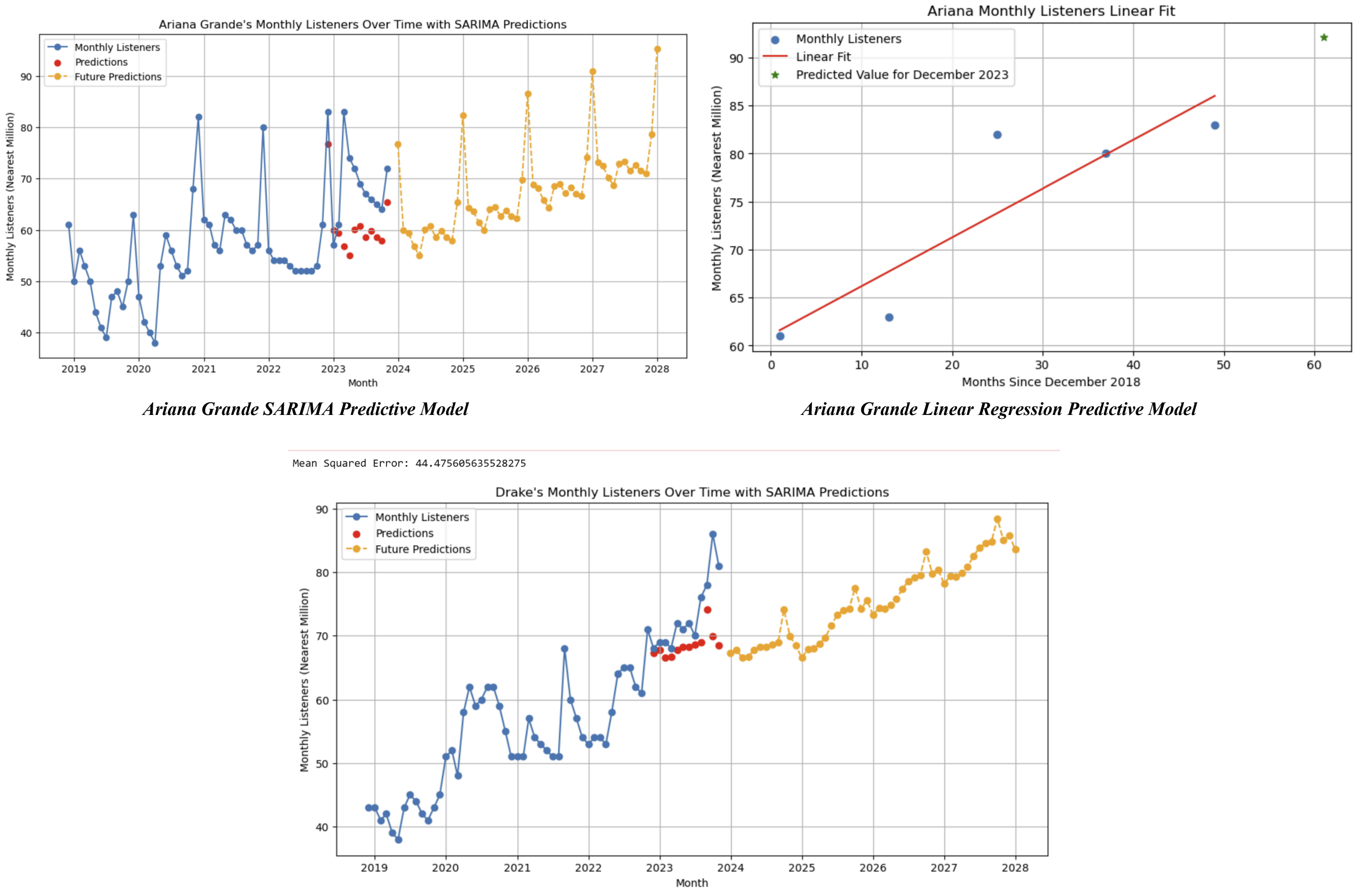
In Figure D on the bottom left, Drake's top 10 songs list we can see that Drake has no Christmas songs in his top 10 as Drake is not known for Christmas music. Drake has been extremely active lately releasing an album in 2018, going on tour in 2019, releasing another album in 2021, releasing 2 more in 2022, another in 2023, and then going on tour again in 2023.

In Figure D on the top right, the dot plot of Ariana Grande's monthly listeners from 2018-2019, we see several things. The first thing worth noting is that there are peaks in her average monthly listeners in the month of december each year before there is a drop off to January the next month. Unlike the Christmas giants however, Ariana's peaks and drop offs are much less extreme due to the fact that she was making live appearances and releasing new music along with the fact that the majority of her music is not Christmas related. We can also notice that unlike the Christmas giants, Ariana has a generally upward trend. Also in contrast with the Christmas giants, Ariana did not have a stagnant average monthly listeners during the non-holiday months of the year.

When looking at Figure D on the bottom right, the dot plot of Drake's monthly listeners from 2018-2023, we notice that there is consistent growth with very few extreme spikes in the graph. This is very different from the Christmas giants being that Drake's listeners only stagnate or start to decrease when he has not been active recently and he does not simply drop listener numbers drastically when the Christmas season ends.

As previously mentioned Ariana was extremely active between 2018-2021 which would explain why there are multiple spikes throughout the non-holiday months of that 3 year span. That 3 year span also includes the majority of her steady upward trend which is most likely the result of her activity increasing her popularity as well as spotify having a large increase in popularity. The year of 2022 saw a slight but steady decline in Ariana's monthly listeners due to her inactivity followed by her all time peak in December of that year. This came as a result of her inactivity over the course of that year followed by the Christmas spike.

Similarly to Ariana, Drake's listener numbers are highly related to his activity in releases and appearances. Those events are what has been driving his growth along with the natural growth of Spotify.



Ariana Grande SARIMA Predictive Model

Ariana Grande Linear Regression Predictive Model

```
Mean Squared Error: 44.475605635528275
```

Drake SARIMA Predictive Model

xFigure E. Predictive Models of Ariana Grande and Drake

In Figure E in the top left, the SARIMA model for predicting Ariana Grande's Monthly Listeners, we trained the model using the data from 2019-2022 to predict the year of 2023 data and then get a mean squared error value. The mean squared error value was approximately 128.44 million listeners. This is an extremely high error value considering that Ariana Grande has not once eclipsed even 90 million listeners. This would lead us to believe that the SARIMA prediction model for Ariana Grande is significantly less accurate than the SARIMA prediction models for the Christmas giants.

In Figure E in the top right, the linear regression model for Ariana Grande's peak monthly listeners for the month of December each year, we see a prediction of 92.10 million monthly listeners for the 2023 December peak. We would not expect this prediction to be as accurate as the linear regression models for the Christmas giants. The reason for this is that the Christmas giants' peaks and drop-offs are mainly related to Christmas time. Ariana on the other hand, has less severe Christmas peaks, but her non-holiday months are much more reliant on her activity, which in turn will also have some impact on the Christmas peaks because there will be people who continue to listen to her regardless of Christmas. Therefore this prediction model cannot be as accurate as it was for the Christmas giants because we do not have a model to predict Ariana's activity meaning we cannot accurately predict her monthly listener trends.

In Figure 5 on the bottom we have Drake's SARIMA predictive model. Going through the same process of training the model, predicting the recent year, and then calculating the mean squared error value of approximately 44.5 million listeners. This value is very large considering that 44.5 million would be more than half of Drake's total listeners at any point on this graph, however that is nearly 3 times lower than the mean squared error value for Ariana's prediction. This comes as a result of Drake being more consistent with releases and appearances and not having variable spiking during holiday time periods. This error value is however roughly 10 times greater than the error in Mariah Carey's SARIMA prediction. This is again a result of Drake's data being dependent on activity and not seasonality.

It is important to note that as the years go on, Spotify is getting more popular and has more people using the streaming service. This can be seen as the peaks of the linear regression models are rising throughout the years and the "relatively low" number of monthly listeners for the Christmas Giants get larger as the years go on. However, with the increasing popularity of Spotify, we still determine the peaks of the Christmas Giants are correlated with the time of the year.

4. Conclusion

By analyzing data visualizations and using both SARIMA and linear regression predictive modeling, we were able to come to a conclusion that artists with popular Christmas songs have a strong increase in monthly listeners during Christmas time. On the other hand we also noticed that the otherwise popular artists not known for Christmas music saw strong increases in monthly listeners more closely tied with their new album releases and tours. There are numerous ways to continue building on these results including using past instances of releases and touring to attempt to predict the schedule of artists albums and tours in the future. The time series analysis process could also be used for other holidays like Easter or Halloween to see if the tie between holiday and artist stands for those other artists. These results aid in the gathering of information for marketers in this field. This could include record labels, advertisers, artists, and many other corporations who might seek to have a collaborative commercial or appearance with certain artists. For instance, it may not be in Mariah Carey's favor to drop a non-Christmas single in November, as that is when the Holiday spike starts and the song would most likely be overlooked, however releasing that music during December would have much better results. Doing a commercial with Mariah Carey for a Christmas themed product would probably be more effective than doing that same commercial with an artist like Drake who is not seen connecting with Christmas as much. Overall, our project shows that anyone can use a topic they love, like music, and answer a question or gain real insights from that topic with the aid of data science.

5. Roles*

Humphrey: Initial Idea, Linear Regression, SARIMA Modeling, Abstract, Results/Discussion of Christmas Giants

Josh: Time Series Analysis, Gathering Impacts of Significant Events, Introduction, Methods

Shane: Data Collection, Spotify API, Web Scraping of Sources, Results/Discussion of Other Popular Artists, References

*- The roles are not exhaustive; the project was very collaborative in regards to the code

References

[1]"Ariana Grande Concert & Tour History: Concert Archives." *Ariana Grande Concert & Tour History | Concert Archives*, www.concertarchives.org/bands/ariana-grande. Accessed 20 Dec. 2023.

[2]"Ariana Grande Albums and Discography." *Last.Fm*, www.last.fm/music/Ariana+Grande/+albums. Accessed 20 Dec. 2023.

[3]"Drake Albums and Discography." *Last.Fm*, www.last.fm/music/Drake/+albums. Accessed 20 Dec. 2023.

[4]"Drake Concert & Tour History (Updated for 2023 - 2024): Concert Archives." *Drake Concert & Tour History (Updated for 2023 - 2024) | Concert Archives*, www.concertarchives.org/bands/drake. Accessed 20 Dec. 2023.

[5]"Mariah Carey Concert & Tour History (Updated for 2023): Concert Archives." *Mariah Carey Concert & Tour History (Updated for 2023) | Concert Archives*, www.concertarchives.org/bands/mariah-carey. Accessed 20 Dec. 2023.

[6]"Mariah Carey Albums and Discography." *Last.Fm*, www.last.fm/music/Mariah+Carey/+albums. Accessed 20 Dec. 2023.

[7]“Spotify Stats.” *Spotify Stats*, spotifystats.com/. Accessed 20 Dec. 2023.

[8]“Wham! Concert & Tour History: Concert Archives.” *Wham! Concert & Tour History | Concert Archives*, www.concertarchives.org/bands/wham. Accessed 20 Dec. 2023.

[9]“Wham! Albums and Discography.” *Last.Fm*, www.last.fm/music/Wham!/+albums. Accessed 20 Dec. 2023.