SONG SALE FORECASTING BASED ON STANDALONE MUSIC STREAMING A CASE STUDY OF SPOTIFY ON POPULAR U.S. CHART SONGS

AMAYO OKI OTAMBO
101667

Submitted in partial fulfillment of the requirements for the Degree of Bachelor of Business Science Finance at Strathmore University

Strathmore Institute of Mathematical Sciences
Strathmore University
Nairobi, Kenya

February 2021

DECLARATION

I declare that this work has not been previously submitted and approved for the award of a degree by this or any other University. To the best of my knowledge and belief, the Research Project contains no material previously published or written by another person except where due reference is made in the Research Project itself.

Amayo Oki Otambo

10/02/2021

This Research Project has been submitted for examination with my approval as the Supervisor.

Dr. Linda Chaba

10/02/2021

Strathmore Institute of Mathematical Sciences
Strathmore University

ABSTRACT

The recent rise in music streaming services and audiovisual platforms has facilitated listeners and music enthusiasts from all around the globe with even easier and cheaper access to music than ever before. However, this technological footprint has also had a significant impact on the way music is marketed and distributed, thus affecting music revenues, record sales and royalty rates. Nevertheless, an ongoing debate seems to exist as to whether music streaming actually substitutes or complements record sales, with various research papers significantly backing each view. This project seeks to add on to the existing literature by looking at a more recent case study of Spotify in the United States. The plausible impact of streaming services on digital record sales is thus investigated using weekly data from June 2019 to June 2020, comprising the United States' Top 200 most streamed songs from Spotify as well as the Top 100 most sold songs in the U.S. as observed by Rolling Stones. These variables are then incorporated into two independent univariate time series models to account for time-lag effects that may exist in each, as well as investigate the plausible influence of time on both variables. The causation effects between the two variables and their lagged forms are then subjected to a Granger-Causality test using a vector autoregressive model to investigate if there is in fact a relationship between them. Ultimately, the findings exhibited in this study seem to conclude that there is no relationship between streaming numbers and their equivalent record sales. Furthermore, both variables seem to exhibit random walk properties as both the time series models generated in this study do not seem to adequately describe the song sales and streaming paths, as demonstrated by the sample, suggesting that future values of both variables cannot accurately be predicted. From this, it is recommended that future research could instead try to focus on the effects of the two based on long-term data and instead focus on proving whether both the streaming numbers and song sale units do in fact follow a random walk process.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

1.1. Background

Music: the cultural phenomenon, commonly described as an art form, whose primary medium is sound. According to Wilkinson (2000), one of the largest and earliest collections of musical instruments was believed to date back to as early as 7000 BC. Since then, music has continued to evolve and has proved itself to be a fundamental element in people's lives. More so, with technological advancements continuously driving innovation, music has taken up many different forms and more-personalized genres and subgenres as well as become a more significant part of different fields and disciplines.

1.1.1. The Gramophone

Music production and distribution can be traced back to the early 1910's when the gramophone was invented. According to Orpheus (2017), the gramophone was considered the first consumer playback system that was primarily used to distribute printed music scores that buyers could use to play popular songs on the piano. It is from this that the term *royalties* were abstracted and would in later years be modified to what we know as the sums paid to recording artists today.

1.1.2. The Age of Vinyls, Cassettes & the Introduction of CDs

The early 1980's saw vinyl and cassette sales reach an all-time high, before Sony & Phillips sealed the era with the introduction of CDs. Record labels, the intermediaries between artists and consumers, at the time seemed to greatly prefer CDs to vinyls as they were lighter and smaller in size and could thus be stored and shipped much more efficiently - due to the apparent reduction in transportation costs. They also later realized that, despite charging similar prices, the demand for CDs was much higher than that of vinyls and could thus profit =more by selling CDs at even higher prices while simultaneously keeping the royalties that artists earned constant.

1.1.3. The Piracy Invasion

However, unlike vinyls, the data stored in CDs could be copied and reproduced infinitely without distorting the quality of the data. In the early 1990's, record labels were then faced with the risk of CD piracy that saw the sale of illegally-copied yet same-quality CDs being sold at much lower prices. More so, the introduction of MP3 formatting in 1995 facilitated the leeway to peer-to-peer file sharing. With the commercialization of the Internet in the United States at around the same time, file sharing sites were soon established and allowed users to share digital files in MP3 format either for free or significantly lower prices. Record labels, in their attempts to recover the situation, tried to enforce copyright laws that pressured the governments to arrest pirates and sue several file sharing sites. This overall had an adverse effect on consumers as it isolated the music enthusiasts who hitherto were seemingly their best customers.

1.1.4. The Age of Digital Downloads

In 2001, piracy was reaching an all-time high as approximately 40% of all sold music CDs were pirated disks. With an initial offer by Steve Jobs to distribute singles and albums as two separate entities instead of a part of one record, many record labels seemed to find the idea unsettling. This was further propelled by the idea of selling a song for $0.99 (US dollars), which hitherto had never been done before. However, succumbed by the circumstances, numerous record labels were eventually enlisted, and iTunes officially came to be in early 2001. By 2003, Apple was able to sell 25 million digital downloads, taking a 30% portion on every sale. This jump was further propelled by the marketing strategy that Apple had set in place, putting convenience at its forefront.

1.1.5. The Start of Spotify

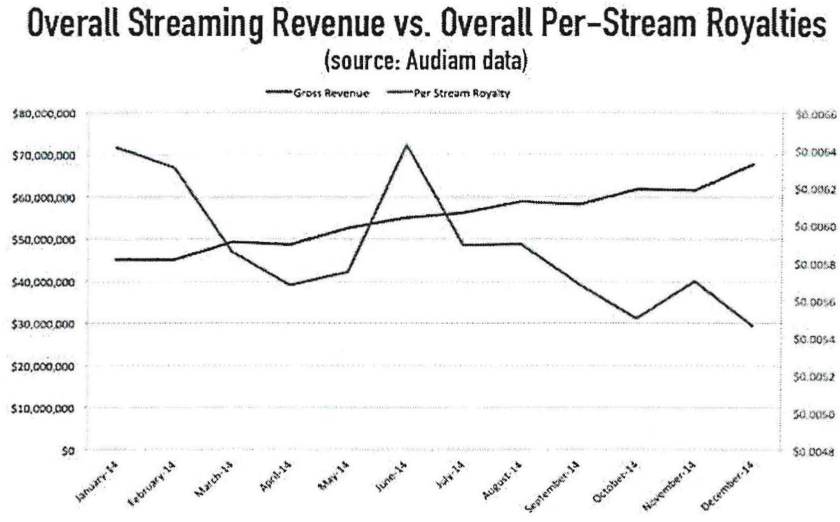
With the idea that what consumers really wanted was access to music rather than the physical records themselves, Daniel Ek (the then CEO of uTorrent) co-founded Spotify in 2006. Through Spotify, he offered to pay record labels a licensing fee and guaranteed cash advances in exchange for their catalogues of music. These catalogues were then made available on the platform and thus to a consumer after paying a periodical subscription fee. Moreover, the record labels were made shareholders in the company, and artists would, in turn, earn from a smaller per-stream rate. This served as the beginning of what would be the turnover of the industry's entire business approach.

1.1.6. The Streaming Era

With the rise of similar streaming services and audiovisual platforms such as YouTube, people have been facilitated with even easier and cheaper access to music than ever before. According to the Recording Industry Association of America (RIAA, 2019), as of 2019, nearly 80% of music revenues in the United States comes from streaming of music accounts alone. This data encompasses both ad-supported and subscription-based streaming services. However, this technological footprint has also had a significant impact on the way music is sold.

By the start of 2013, record labels, that were formerly the kingpins of the industry, were forced to develop more innovative ways of generating revenue besides physical record sales. Gamal (2012) posits that this was likely due to the further reduction in production costs as music equipment has gotten less and less expensive and more widely available overtime. More so, in an era where social media was becoming more and more prevalent in people's lives, the costs of marketing music have nearly been decimated by open and free web communication. Nevertheless, it is evident that record labels still play a big part in linking artists to the public, and more so, in linking artists to their income as most record labels get to control the rate of return from which governed artists earn their share.

Figure 1: A line graph depicting the US Overall Streaming Revenue and Overall Per-Stream Royalties for the Year 2014



Source: Orpheus, R. (2017). The Economics of the Music Industry and the Impact of Digital Technology. Retrieved from https://medium.com/@rodneyorpheus/the-digital-music-industry-an-overview-965915ae1fab

However, these royalties are not consistent but change with regards to the volatilities of the streaming service being supported. As displayed above, the per-stream royalty rate seems to exhibit an erratic but overall downward trend despite a steady increase in gross revenue earned from streaming services. Orpheus (2017) suggests that this may, in part, be due to the operational costs behind them. An example is Spotify's freemium model that is ad-supported hence allows users to have free access to the catalogues of music provided by record labels. Forced to cover their costs, streaming services such as Spotify often resort to decreasing the per-stream royalty rate and keep the record label cash advances constant to ensure that their consumers have continuous music support.

1.2. Problem Statement

Several studies overall have revealed that there exists a continuous debate as to whether the digitization of the music industry, especially in regard to the most recent introduction of streaming services, has had a positive or negative effect on the revenue earned by the music industry. Despite the purported decrease in per-stream royalties, a recent study by Lee et. al. (2016) seems to postulate that the increase in popularity amongst streaming services has led to an equivalent increase in album sales. This overall seems to have a large contrast to the digital download age – where numerous consumers fell victim to online pirated music, thus making record labels profit-less. According to Peitz & Waelbroeck (2004), the format provided by digital downloads only seemed to heighten music piracy through peer-to-peer file sharing sites such as Napster, which overall had an unfavorable effect on the music industry.

However, with the introduction of streaming and evolution of the Internet (and thus Internet-supporting devices), consumers seem to be more encouraged than ever before with even easier access to music services crowning the music market. According to RIAA's Music Revenues Report (2019), paid subscriptions to on-demand streaming services seemed to contribute the largest share of revenues for the U.S. music industry in 2019, totaling to about \$6.8Million, thus accounting for about 61% of the total recorded music revenue. The number of paid subscriptions grew by 29% from the year prior, reaching a new average of 60.4Million subscribers. This is nearly a 460% increase from the number of subscriptions totaled in 2015. On the other hand, ad-supported streaming grew by about 20% from the year prior, totaling to about \$908Million. However, since these services were primarily ad-supported, they contributed only 8% of the total U.S. music revenues for the year.

Whilst the number of subscriptions to streaming services seem to be on a generally uphill climb, other forms of music revenue seem to be declining overtime. According to RIAA's Music Revenues Report (2019), "the total revenues from digitally downloaded music were down 18% to \$856Million in 2019, marking the first time since 2006 that revenues from downloaded songs and albums fell below \$1Billion." Overall, digitally downloaded music contributed to only 8% of the U.S. music revenues in 2019, marking a 63% decrease from the \$2,314Million earned in 2015. Moreover, revenues from physical music products decreased by 0.6%, from 2018, to \$1.15Billion in 2019. Bukszpan (2019) proposed that the fall would have been worse had it not been for the trending rise in vinyl purchases. As a result, revenues from vinyl purchases increased by 19%, from 2018, totaling to \$504Million in 2019. This, in turn, offset the otherwise declining revenue from CD sales, which decreased by 12% from 2018, totaling to about \$615 Million in 2019. While artists and record labels might remain hopeful of the steady rise of vinyl sales, they only accounted for 4.5% of the total U.S. music revenues in 2019.

1.3. Research Objectives

Lee et. al. (2016) seems to presume that physical album sales have increased since 2007 – around the same time Spotify was started – with various albums exceeding 100,000 sales. However, this study considered only Korean music data and thus focused solely on the Korean music industry, giving rise to the question if the same would be applicable in other major music industries around the globe. Furthermore, it is plausible to question if one can develop predictive models to predict record sales and streaming numbers as well as determine the influence of streaming services on record sales - that is, whether there is a statistically significant causal relationship between the two. This project thus aims to investigate this with regards to the current economic environment governing the music industry. This study shall build up on the arguments set out by Lee et. al. (2016) that presupposes a positive effect of online music streaming on album sales.

Specific objectives

1. To determine the behaviour of streaming data overtime
2. To determine the behaviour of record sales overtime
3. Using the models developed in Chapter 4 to forecast future record sales and streaming numbers.
4. To investigate the plausible causal relationship between standalone streaming services and record sales.

Research questions

1. How does streaming data behave overtime?
2. How do record sales behave overtime?
3. Is there a causal relationship between standalone streaming and record sales?

1.4. Significance

This study aims to aid record labels in seeing the importance of the impact of streaming services in the music industry and thus further invest or lessen the restrictions on streaming services' access to the equivalent catalogues of music. Alternatively, this study may encourage music artists to seek other means of generating revenue besides per-stream royalties that they can thus capitalize on. Furthermore, this project aims to contribute to the existing literature and try to fill in some of the gaps made by previous research.

CHAPTER 2: LITERATURE REVIEW

2.1. Theoretical Literature

2.1.1. The Terms and Types of Music Streaming Services

Generally, music streaming services can be categorized into two main types: standalone streaming services and value-adding streaming services. Standalone streaming services involve the likes of Spotify, SoundCloud, Deezer and Tidal that operate as an entirely independent business entity; whereas value-adding streaming services include services whose operations are a subgroup or subsidiary of an even larger business or company. An example of a value-adding music streaming service is Apple Music which, according to Orpheus (2017), is a value-adding approach to the Apple ecosystem. He further concludes that Apple Music helps ensure that consumers remain linked to Apple's ecosystem and are thus propelled to invest more in Apple products. In the latter case, value-adding services like Apple Music thus enhance the consumer experience that Apple has centered around its products, hence giving them a greater perception of value. Whereas value-adding services' main objective is to add value to an otherwise existing product or service, standalone streaming services, like most businesses, aim towards earning profits.

Standalone streaming services can further be divided into two forms: paid subscription services and ad-supported services. Whereas the former allows consumers to access a grand catalogue of music by paying a subscription fee at the start of every pre-set period (typically monthly), the latter is usually free and depends on revenue from advertising. Such advertisements often play out at some point in the listener's experience and prompt listeners to invest in other goods, services or businesses that may not necessarily be related to the streaming service. In some cases, listeners do not even need to subscribe to access music from an ad-supported streaming service. Examples of such services include Deezer and YouTube.

In other cases, users are given the option of choosing the version they prefer from the same streaming service. An example of this is Spotify, whose consumers are given the option of choosing between the unlimited ad-supported "freemium" version and the subscription-based ad-free premium version. In Spotify's case however, the freemium version only supports shuffle-only playback. This means that users are exposed to scheduled and broadcasted music, and do not have the on-demand (interactive) access that their paying counterparts have. However, as Desai (2019) seems to propose, the "freemium" model is essentially meant to drive consumers towards the premium subscriptions by initially providing users with a mimicked experience of the service and its benefits. He further suggests that Spotify sets out to create a sustainable habit amongst its consumers, with an aim to encourage freemium consumers into using the more premium version. This is further enhanced by the already-low subscription fees, as compared to the price of purchasing individual albums physically or digitally.

2.1.2. The Sampling Effect Vs. The Substitution Effect

As highlighted in an article by Goldman (2010), by 2009, the total revenue from U.S. music sales and licensing dropped to $6.3Billion from $14.6Billion in 1999. This was predicted to decrease even more as the years progressed. In their attempts to revive the industry, record labels and artists attempted to offset the decline in record sales by investing in digital music products. A large pool of literature discussing the effects of digitizing the music industry is thus available.

2.1.2.1. The Introduction to the Sampling and Substitution Effects

Peitz and Waelbroeck (2004) seem to suggest that legitimate record sales can be supplemented by digitizing such records. They did so by implementing a multi-product monopoly model where consumers were introduced to a Salop circle whose products were ranked as superior or inferior regarding both their originals and copies. By looking at the relationship between unit demand and full consumer participation, they were able to obtain sustainable results from the study. They seemed to agree that the sampling effect, as purported in their model, enables consumers to weigh in and find better matches to their tastes, which in turn leads to high file-sharing profits. However, they further stipulate that such can only happen theoretically, given that the original version of the record and its digital copy are differentiated enough, but offer the willing listener an ideally similar experience. This hypothesis thus holds only if there is a sufficient balance between consumer taste heterogeneity and product diversity. Therefore, the sampling effect is only applicable to the listeners whose tastes ideally are matched by the record. This means that listeners are likely to buy a physical record of a song or album they have consumed digitally, only if they like it. Otherwise, a consumer's willingness to pay for legitimate digital downloads or physical records reduces if either aspect is not present.

Whereas earlier studies centered around the sampling effect, the impact of music piracy revealed the likely substitution effect, where consumers often substitute physical records for their digital counterparts. Numerous academic studies have thus debated on the impact of music piracy on record sales. Using household data from the Consumer Expenditure Survey (CEX), Michel (2006) appears to prove that online file-sharing technology seemed to have led to the greater part of the downfall of record sales between 1999 and 2003. By measuring the change of computer ownership against the changes in household expenditures on music, using data from 99 American cities, Michel seems to conclude that the relationship between computer ownership and music purchases drastically weakened after online file sharing became a feasible option for consumers. This is especially for heavy music consumers who, at the time, opted to obtain large amounts of music through the free and easy file sharing platforms, thus confirming the substitution effect. This is further emphasized by Liebowitz (2005), who seems to blame the significant reduction of record sales on increased Internet penetration, which in turn led to the further development of free peer-to-peer file sharing platforms such as Napster. He further suggests that per capita record sales would have been higher had file sharing not impacted the music industry, as it did. Furthermore, the exposure of consumers to such file sharing platforms results in evolution of the substitution effect, caused by the reduction of the probability of actually buying music, and thus resulting in the gradual decline of record sales (by as much as 13% in the examined period).

Both studies seemed to emphasize the substitution effect as neither study showed evidence of an increase in record sales during the examined periods. Nevertheless, in both cases, the periods examined were deemed too short to investigate any long-term effect of music piracy on record sales. Both papers seemed to look only into the 3 years immediately after the inception of Napster. This results in a short-term perspective of the impact of music piracy on record sales, which may differ from the long-term effect of the same.

2.1.2.2. The existing variations of the sampling and substitution effect

There has been similar research done on the impact of the effect of Internet penetration on the sales of digital media content. Smith and Telang (2009) seem to suggest that movie broadcasts on free-to-air networks results in a notable increase in DVD's sales. They study this by looking into the effect of movie broadcasts on the demand on DVD's. Using Amazon.com as the proxy for a legitimate DVD provider and BitTorrent as the proxy for the pirated channel, they conducted the study with data, which included all movies that were shown over-the-air and on cable television during an eight-month period (between 2005 and 2006). Their findings revealed that movies shown on free-to-air stations did increase the demand of DVD's. However, not only did the legitimate free distribution of movies, through the free-to-air channels, increase the sales of DVDs at Amazon.com, but also increased the illegal downloads for the movies available on BitTorrent during the broadcast period. This shows the possible co-existence of both the substitution and sampling effect that seemed fairly balanced out during the broadcasted period. Nevertheless, they conclude that their findings suggest that giving away free content on one channel (the free-to-air channel) can stimulate sales in a paid channel. However, this is only possible if the free and paid products are sufficiently differentiated, and appeal to the customers willing to buy the DVD. For this, the content thus needs to appeal to the available customer segment to minimize illegal downloads and harming paid sales. The findings in this study seem to go hand in hand with Peitz and Waelbroeck's (2004) study whose conclusions seemed dependent on a customer's taste and their overall utility with regards to the content at hand. However, the study only accounted for sales rank data rather than actual sales data.

Since the global takeover of streaming services, there have been a few studies conducted on the actual impact of legal streaming services on the music industry. Nguyen, Dejean and Moreau (2013) conducted a study to determine whether the consumption of music through streaming services is a substitute or complement to physical music consumption modes. Based on a survey of 2,007 Internet users, representative of the French & Brittany population, they distinguish their sample behaviours with regards to a CD variable (how many CDs the consumer purchased in the last 12 months), and a live concert variable (how many live concerts the consumer has attended over the same period) – both of which act as the dependent variable. For the independent variable, a dummy variable, representative of the streaming habits of consumers, is used. The dummy takes the value of 1 for consumers that declare that they typically acquire their music from online streaming services such as YouTube, Spotify, Deezer amongst others. By further including variables that account for consumer-to-consumer promotion, consumers' radio habits, consumers' TV habits, their reading habits with regards to music articles, and their music taste (where a consumer is attached as strong, moderate and weak with regard to their mode and frequency of music consumption), they established a tobit regression with CD sales and live concert attendance as the independent variables. With this model, they seemed to conclude that streaming has no effect on offline music sales but has a significant positive effect on attendance of concerts, more so by international artists rather than local artists. According to the study, streaming thus acts as a weak substitute for recorded music and may more likely play a role as a discovery tool rather than an alternative for recorded music. This overall seems to neither completely confirm nor deny the claim that such services may have a sampling effect on actual recorded music. However, their study seemed to only account for CD sales, which makes up just 80% of the French music market, at the time of the study. Moreover, the use of the dummy variable restricts the implementation of a more robust analysis as it does not consider the magnitude of per-pay downloads. On top of that, the instrument variable used for consumers' reading habits regarding music articles, seemed to be unsatisfactory as much as online and offline music sales are concerned. Nevertheless, using a probit regression, they seem to suggest a complementary effect between streaming services and digital music sales, and justifies further research on this topic.

Aguiar (2017) examines the effect of a listening cap on free versions of streaming services on the music purchasing and piracy behaviour of consumers. Using Deezer as its free and mobile-restricted streaming service, sample data consisting of 5,000 voluntary French individuals was taken over a one-year period starting from January 1, 2011 to December 31, 2011. Each person's type of subscription was identified and recorded both before and after the listening cap was implemented by Deezer on June 6, 2011. Out of the sample data, Aguiar seemed to note an increase by approximately 24% of people who had elevated their subscription to the more premium service after the cap was introduced. For the study, the number of clicks to the set of alternative music consumption websites is set as the dependent variable, and the dummy variables Deezer (which takes up the value 1 if a consumer visited Deezer before the listening cap was imposed) and Cap (which takes up the value 1 if a consumer continued using the free version of Deezer after the cap was imposed) are set as the independent variables. Moreover, a control variable is set as an independent variable to control for the variation in alternative consumption across individuals, which is constant overtime, which includes the time-invariant differences in music tastes of each person in the sample. Also, another control variable is introduced as an independent variable to account for the alternative consumption that is common to all the individuals in the sample. Overall, Aguiar seems to establish that the implementation of the listening cap led to a negative effect on the visits to both licensed and unlicensed music download websites. Furthermore, Aguiar appears to suggest that the Deezer users who felt constrained by the listening cap visited such websites 2% less than they would have had the cap not been introduced. This proposes that free streaming poses a positive effect on alternative music consumption, further supporting Nguyen, Dejean and Moreau's (2013) claim that streaming services can act as a discovery tool and lead to further stimulation of alternative music consumption, that offers more mobility. However, the study seemed to account only for free streaming services or free versions of streaming services. It hence does not necessarily apply the same results for premium streaming services that instead offer full mobility to consumers.

2.2. Empirical Literature

Aguiar and Waldfogel (2015) examine the effect of interactive streaming on the revenue earned for the recorded music industry. In their study, they weigh in on the perspective that "streaming services enable sellers to engage in bundling with the promise of increasing revenues, profits and consumer surplus". This bundling, if successful, would cover the unpaid consumption and deadweight loss, obtained by music that does not generate revenue through individual track sales, by interpreting this loss as the willingness to pay for the bundled offering. Using the top 50 stream data from Spotify between 2013 and 2015 and weekly data on digital sales for 21 countries, Aguiar and Waldfogel (2015) employ a multivariable regression, for their aggregate approach, that tries to uncover the causal impact of streaming on song sales across 21 countries:

$$Qcts = Yc\,Yt + \alpha Sct + \varepsilon ct$$

where $Qcts$ represents the total consumption of song s in country c during week t, Sct represents the equivalent Number of Spotify streams for country c in year t, Yc and Yt represent the country-fixed and weekly-fixed effects respectively, and εct represents the error term. In this model, the sales of a particular song in a particular country in a particular week act as the dependent variable, and the analogous streams of a particular song in a particular country in a particular week acts as the independent variable. The model also controls for the country fixed effect and the song-specific time pattern by including two new independent variables that represent both effects respectively on the dependent variable.

They appear to discover that a positive relationship does exist between streaming and track sales on a song-level basis. However, at the aggregate level, a negative relationship between interactive streaming and track sales is noted. From this, they seem to suggest that "Spotify is better viewed as a form of bundled sales than as a promotional channel", thus backing up the substitution effect. However, there seems to be sufficient evidence to show that Spotify does displace piracy as the new revenue generated through streaming payments may offset the revenue reductions from the illegal sale of digital downloads.

Whereas the model seems to offer a considerable approach to the investigation on the plausible displacement of song sales by streaming, the model does not consider the effects of lagged streams and lagged song sales, thus ignoring the time-variant effects of both variables. Furthermore, with the continuous increase in streaming subscriptions, it seems that this may result in a rise in the overall revenue, even in cases where streaming may displace individual track sales. However, this is only possible if the streaming payment is large enough to offset the extent of sales displacement. Despite this positive outcome, the study does not account for the effect of non-interactive streaming services on the overall music sales, at an aggregate level, hence possibly losing out on a significant part of the current music industry.

In a study conducted by Lee et. al. (2016), the impact of online streaming services on record sales is investigated to determine the effect of streaming on physical music products. Using monthly collections of record sales and streaming data consisting of the top 200 streaming and sales figures from the official Korean Gaon Music Chart between March 2011 and July 2013, a two-stage least squares regression is built, in which the number of monthly record sales [ln(sales)] is the dependent variable, and the number of monthly music streams [ln(num_streaming)] is an independent variable:

$$\ln(sales_i) = \beta_1 \ln(num\_streaming_i) + \beta_2 \ln(previous\_sales_i) + \beta_3\, rating_i + \beta_4\, num\_song_i + \beta_5\, price_i + \beta_6\, gender_i + \beta_7\, member_i + \varepsilon_i$$

Moreover, the model consisted of album-specific and artist-specific characteristics, that consist of the number of television promotions each month, minimum price per music record (price), number of songs included in one record (num_song), an audience rating (rating), gender of the artist (gender), number of members in a group (member) and the entertainment agency that represent the artist. By controlling for time lag nature of listeners' interests on released music (using the "previous_sales" variable), as well as the interaction between the number of promotion appearances and the size of the entertainment agency, the results obtained from the model seems to suggest that online music streaming has a significantly positive effect on music record sales. They thus conclude that online streaming services promote offline record sales.

This relationship seems to stand regardless of the price or rating of the song or album. However, the number of promotional appearances seem to increase the rate of streaming, further confirming Smith and Telang's (2009) view that giving away content freely on the free-to-air channel can stimulate sales in a paid channel; ultimately, confirming the sampling effect. However, the data used in this study only accounted for local artists and seemed to ignore the effect by international artists on the same music industry. Moreover, this study only examined the Korean music industry, thus posing the question if the same results would be maintained in other music markets.

2.3. Research Gap

There are many studies, such as Lee et. al. (2016); Aguiar (2017); and Aguiar and Waldfogel (2015), that focus on the early effect of streaming services on record sales. These papers center around the early 2010's when streaming services were beginning to gain traction. However, none have focused on the more recent effects of the same. Moreover, very few studies have focused on the effects of streaming services on the equivalent digital music sales as a lot of the studies are centered on the effects of streaming on physical records sales. Furthermore, given the current trends as depicted in Chapter 1 of this project, the need for globalization of the topic by Lee et. al. (2016) seems more admissible than ever before. This project thus aims to extend the topic on the relationship between streaming services and record sales, with a focus on more recent streaming and song sales data; thus, aiming to further this study with regards to the current economic environment governing the music industry. Furthermore, this project aims to project the results into the near future, given the assumed sustenance of the current trends. This will be done by employing U.S. streaming and record sales data over the past year and implementing an econometric model that can be built to represent the relationship between the impact of audio streaming services on record sales, with regards to the recent U.S. trends exhibited in the examined period.

CHAPTER 3: METHODOLOGY

3.1. Research Design

Overall, this project implements a quantitative research approach as the observations are to be examined using statistical models, from which the objectives can be proved or disproved. The design is further broken down into descriptive, predictive, and correlational research. The descriptive research part involves the derivation of summary statistics that quantitatively describe each variable, such as the mean, median, standard deviation, kurtosis, and skewness. The predictive research part seeks to add on to the descriptive research part by developing single independent models that speculate future observations based on the available data for the song sales and streaming numbers, respectively. The correlational research part then focuses on investigating the relationship between the two variables since the theoretical nature of the objectives allows for predictive and correlational relationships to be drawn out from collected data.

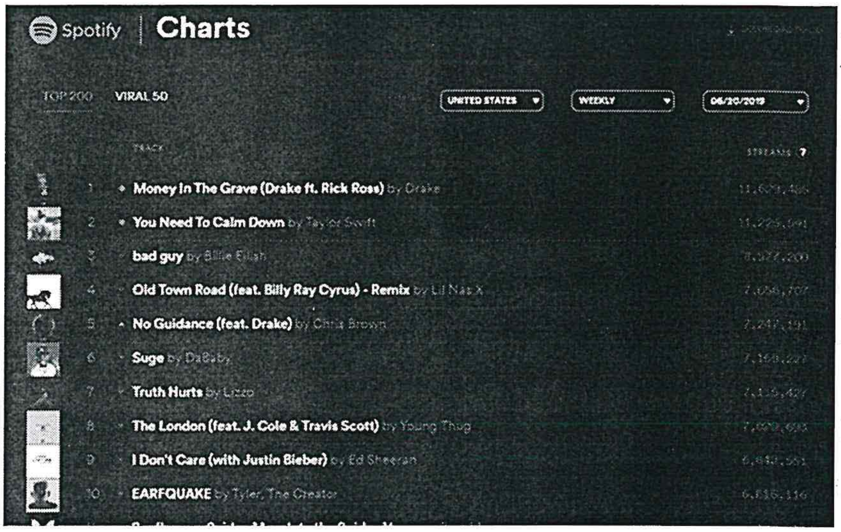
3.2. Data

The data for this study consists of song sales, representative of record sales, and equivalent streaming units obtained overtime. Song units appear to be the most suitable proxy for record sales, as compared to album units as album units are more restrictive and complicated in terms of their data measurability and availability. For example, it would be harder to trace the progress of standout songs, such as album-pre-release songs and hit songs, overtime if album units were used. Therefore, song units seem to be the more acceptable proxy for record sales as they can be observed much easier overtime.

The data for this study comprises weekly U.S. chart data of the Top 200 most streamed songs from Spotify and weekly U.S. rank data of the Top 100 most sold songs, as depicted by the Rolling Stones Music Charts, from June 2019 to June 2020. In either case, Spotify and Rolling Stones were chosen as the sources for the data largely due to their availability and ease of access as opposed to other music streaming services and music charts, respectively. Moreover, as emphasized in an article by The Telegraph (2020), Spotify is widely recognized as one of the biggest, if not well-known, standalone streaming services in the world, whose launch geared the music streaming service market into what it has become today. It thus appears to be the most suitable representation of standalone music streaming services for this study – and thus, the most suitable source for our data. On the other hand, besides the availability of the data, the Rolling Stones Music Charts seem to be the most suitable source for the song sales data because of its systematic similarities with the objectives of this project. This is explained in their General Methodology, where the song chart data only consists of digital song sales and audio stream data – the same two variables that are the main focus of this project – and excludes passive listening, such as digital or conventional terrestrial radio. The Rolling Stones Song Chart is hence the most befitting chart data to use to represent the song sales for this project.

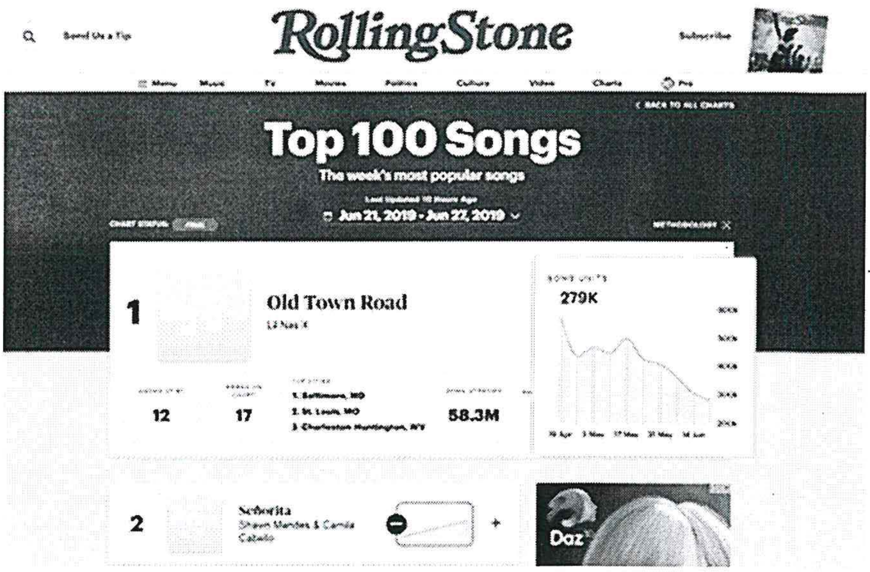
For this study, the weekly U.S. Top 200 most streamed songs are obtained from Spotify's Weekly U.S. Chart (https://spotifycharts.com/regional/us/weekly) as downloadable CSV files, whereas the weekly U.S. Top 100 most sold songs are observed from the Rolling Stones Songs Chart (https://www.rollingstone.com/charts/songs). However, since the Rolling Stones Songs Chart does not provide downloadable versions of their data, the song sales are manually put into their equivalent CSV formats for cleaning and analysis purposes.

Figure 2: A screenshot of the weekly U.S. Spotify Song Streams Chart showing the Top 200 most streamed songs of the week.



Source: Spotify Charts. Retrieved from https://spotifycharts.com/regional/us/weekly.

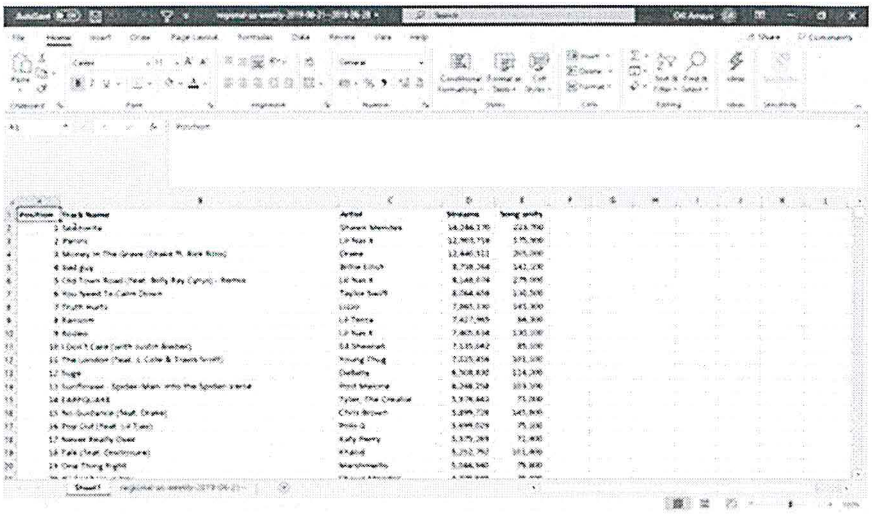
Figure 3: A screenshot of the weekly U.S. Rolling Stones Songs Chart showing the Top 100 most sold songs of the week.



Source: Rolling Stone. Retrieved from https://www.rollingstone.com/charts/songs/

Because the Rolling Stones Chart only ranks the Top 100 most sold songs, the Top 200 most streamed songs are reduced to cater for this difference. The streaming numbers and song sale numbers are manually matched to the song title to obtain only the songs that had both streaming and song sales data.

Figure 4: A screenshot of an Excel worksheet showing the songs whose streaming and song sales data are successfully matched.



Sources: Spotify Charts. Retrieved from https://spotifycharts.com/regional/us/weekly; Rolling Stone. Retrieved from. https://www.rollingstone.com/charts/songs.

Furthermore, a study done by Aguiar and Waldfogel (2015) assumes that the Top 50 streams index is a suitable proxy measure of the Total Spotify streams index. They further investigate this by aggregating the weekly Top 50 and Top 200 Spotify streams across sample countries over 21 weeks. Their investigation showed that the correlations for the 21-week overlap period comes up as 0.99 or 99%, thus indicating that the Top 50 streams index is a valid proxy of the Total Spotify streams index. Borrowing from this paper, the weekly Top 50 songs, that were successfully matched, are hence used as the sample data for this project.

As specified in the RIAA Digital Single Award and the Gelfand, Rennert & Feldman (GR&F) Certification of Audit Requirements (2016), one permanent digital download is aptly and widely considered as one song unit in the United States, for certification purposes, whereas 150 on-demand audio streams are considered as one song unit. To take this into account, the weekly song streams, as obtained from Spotify's charts, are adjusted by dividing each observation by 150 units, to obtain the new set of stream-equivalents sales that would more accurately describe the streaming dataset.

Since the song sales, as obtained from the Rolling Stones Chart, cater for both digital song sales and audio streams, a new dataset consisting of only pure digital song sales are obtained by deducting the derived audio streams from the total song sales to obtain a reliable set of purely digital song sales. It is these sets of purely digital song sales and the set of stream-equivalent sales that are then used to determine the individual univariate times series models and the ultimate vector autoregressive model that will be used to investigate the relationship between the two variables. Time series models are chosen to account for any time-lag effects that may exist as well as investigate the influence of time on and between variables.

To make the investigation of the plausible relationship easier, the dataset of purely digital song sales is assumed to be held and projected by a single record label. This allows for the possibility of cyclicality when constructing the equivalent univariate and vector autoregressive time series model. The stream-equivalent dataset is already adjusted for the possibility of cyclicality since the data is obtained solely from Spotify's charts.

3.3. Data Analysis

3.3.1. Descriptive Analysis

To better understand the behaviour of both the stream data and song sales data overtime, independent descriptive analyses are initially conducted using R. From these analyses, summary statistics, such as the median, mean, standard deviation, kurtosis, skewness and the sum of the equivalent weekly stream and song sales, are independently derived – which are then used to perform the equivalent predictive analyses of the two variables independently. The results are presented graphically to provide a visual representation of the behaviour of each variable overtime. Furthermore, using manual chart tracing techniques such as tracking and observation, samples of artist and song trend data are derived to better explain the track progress and position variation of independent singles overtime. This is included in the descriptive analysis section of this project.

3.3.2. Predictive Analysis

Independent univariate time series analyses are then formed by developing two separate univariate time series models depicting the influence of past stream and past song sales data on current streams and current song sales, respectively. The autoregressive analyses incorporate predictive modelling to determine whether the song sale and streaming paths can be respectively predicted by analyzing their historical data.

3.3.3. Correlational analysis

Lastly, a correlation analysis is then conducted using a Granger-causality test to determine the relationship (if any) between the two variables. This is done by depicting the influence of past stream and song sales data on current song sales, and past stream and song sales data on current streams, as a bi-directional vector autoregressive model.

3.4. Model Specifications

In contrast to Aguiar and Waldfogel (2015)'s model (highlighted in Chapter 2), this project aims to employ the possible lag effects of the variables, as a distinct factor, while investigating the behaviour and relationship between them. However, the country-fixed effects are excluded since the effects are only being investigated in one country the U.S.

3.4.1 The Independent Univariate Time Series Models for Predictive Analysis

Initially, both the stream-equivalent sales and the digital song sales are independently plotted in R and tested for stationarity by observation. Testing for stationarity is important as the derivation, testing and forecasting tools that are used in time series analysis are dependent on the fact that the dataset is stationary. The datasets are further tested using the Augmented Dickey Fuller (ADF) test, with the presence of a unit root - the dataset being non-stationary - as the null hypothesis.

After ascertaining that the datasets are stationary, their equivalent independent time series models are identified using the Partial Autocorrelation and Autocorrelation functions. This is under the assumption that both models are assumed to be some sort of autoregressive model of the structural forms:

$$St = \beta 0 + \beta 1.St - 1 + \beta 2.St - 2 + \ldots + \beta k.St - k + \varepsilon t$$

$$Dt = \alpha 0 + \alpha 1.Dt - 1 + \alpha 2.Dt - 2 + \ldots + \alpha k.Dt - k + \theta t$$

where St represents the current stream-equivalent sales, $St - k$ represents the lagged stream-equivalent sales, Dt represents the current song sales and $Dt - k$ represents the lagged song sales, with k representative of the number of acceptable lagged periods (weeks) and βk *and* αk as their respective coefficients. This model assumes that maximum likelihood estimates (MLE) are the most efficient and sufficient method for estimating the model coefficients, since it tries to maximize the log likelihood or probability of obtaining the data that is already observed. The optimal lag lengths of the respective models are then ascertained by the Akaike Information Criterion, after which the relevant βcoefficients are estimated through linear regression. An autocorrelations test - the Ljung-Box test - is then conducted on the residuals to determine whether the model orders are statistically sufficient. This test is also important as it serves as a diagnostic check to determine if the models are white noise, that is, no serial correlation exists, and thus the existing β coefficients are efficient. If no serial correlation exists, the models will then be used to forecast stream and song sales data, independently, for an assumed near future.

To determine the forecasting precision, the last known period (week) will be removed from each dataset to form a new dataset, which will then use predictive analysis to forecast the data for the last known period. The actual and predicted data for the last known period will then be plotted and compared to see if the predicted data comes close to the actual data.

3.4.2. The Vector Autoregressive Model for Correlational Analysis

To determine if streaming has an effect on the number of digital songs sold and vice versa, a correlational analysis is conducted using a vector autoregressive model of the structural form:

$$St = \beta 10 + \beta 11.St - 1 + \ldots + \beta 1k.St - k + \alpha 11.Dt - 1 + \ldots + \alpha 1k.Dt - k + \varepsilon 1t$$

$$Dt = \beta 20 + \beta 21.St - 1 + \ldots + \beta 2k.St - k + \alpha 21.Dt - 1 + \ldots + \alpha 21.Dt - k + \varepsilon 2t$$

where Dt represents the current digital song sales, $Dt - k$ represents the lagged digital song sales, St represents the current stream-equivalent sales and $St - k$ represents the lagged stream-equivalent sales, with k representative of the number of acceptable lagged periods (weeks) and βkas their respective coefficients. The model above is used under the assumption that the maximum likelihood estimate is the most sufficient and efficient method for estimating the model coefficients, since it tries to maximize the log likelihood or probability of obtaining the data that is already observed. The optimal lag lengths are thus borrowed from the previous independent univariate times series equations and used to form the bi-directional vector autoregressive model that is then tested for causality. This will be done by subjecting the individual coefficients to a Granger-causality test, assuming that all the variables in the vector autoregressive model are stationary. This allows for the formation of joint hypotheses, with the null hypothesis stating that all $\alpha 11, \alpha 12, \ldots, \alpha 1k$ coefficients and all $\beta 21, \beta 22, \ldots, \beta 2k$ coefficients from the vector autoregressive model are equal to zero. This means that streaming does not Granger-cause digital song sales - that is, the available streaming numbers do not contain sufficient information to help predict the number of digital song sales - and vice versa. The alternative hypothesis, in this case, would state that at least one of coefficients listed above is not equal to zero. This means that the streaming numbers do in fact contain some information to predict digital song sales. If the null hypothesis is rejected, impulse response tests will be conducted to examine the magnitude of this relationship. The results are interpreted in the next chapter.

CHAPTER 4: RESULTS & INTERPRETATION

4.1. Findings based on the Descriptive Analysis.

4.1.1. Trend Analysis

Upon loading the data on R, the relevant measures of central tendency – that is, the weekly means and medians – are derived for both streaming and song units to examine the general progression of the data over the selected period. The total weekly streaming and song units are also charted to observe their behaviour overtime.

Figure 5: A graph showing the Weekly Total Song Units from June 2019 to June 2020 (The mean & median equivalents are shown in Appendix I)

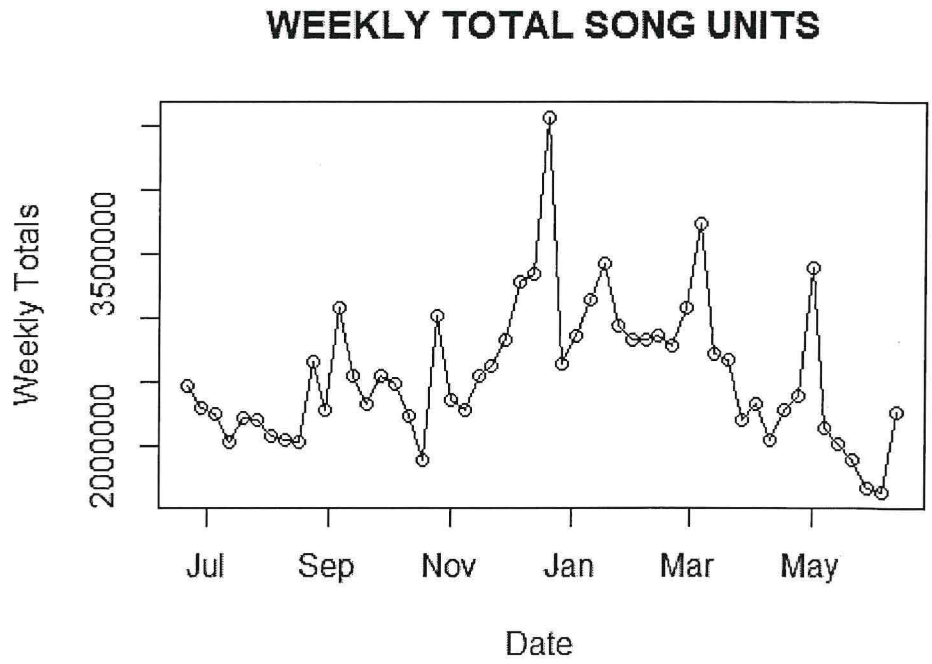
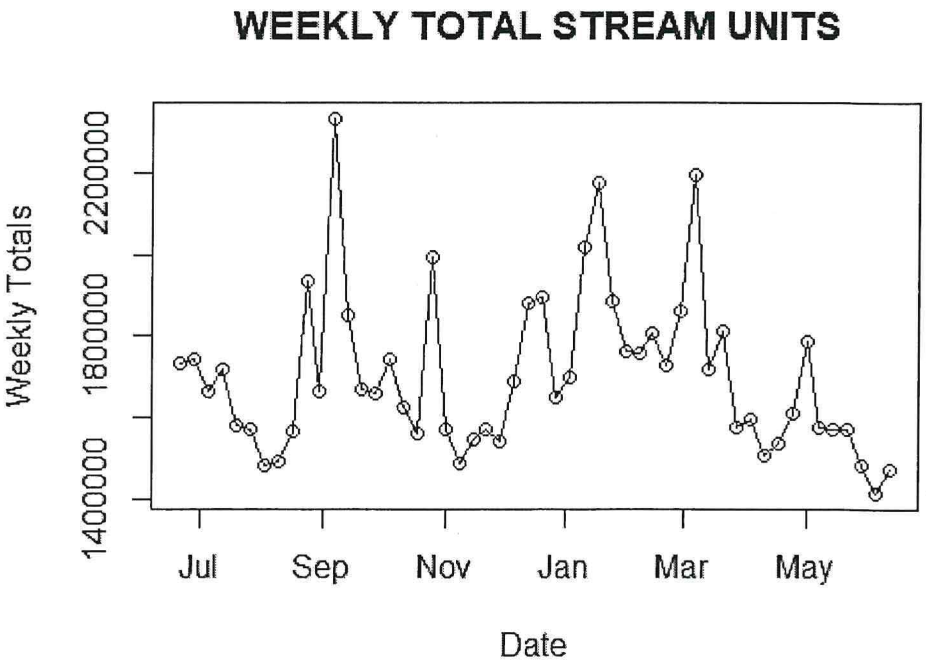


Figure 6: A graph showing Weekly Stream Totals from June 2019 to June 2020 (The mean & median equivalents are shown in Appendix II)



All the above measures seem to vary significantly over time. However, the data on song units (see Appendix I) seems to vary much more sharply than stream data (see Appendix II), with the number of song units reaching extremely high peaks at the end of December 2019 and remaining fairly low at the beginning of June 2020.

On the other hand, the stream data seems to vary less sharply than the equivalent song unit data, with even less sharp peaks and troughs (see Figure 5 and 6). This leads to the presumption that streaming units may contain stronger mean reversion properties than the song units.

4.1.2. Measures of Central Tendency and Variability for the weekly totals

Moving forward with the weekly totals, their various measures of central tendency and variability are derived as follows.

Table 1: A table showing the measures of central tendency and variability for the Weekly Total Streams and Weekly Total Song units

	Weekly Total Song units	Weekly Total Streams
Mean	2,558,029	1,702,839
Median	2,466,734	1,664,910
Standard Deviation	545,266	196,034
Maximum	4,560,228	2,336,776
Minimum	1,634,875	1,412,225
Skewness	1.071033	1.175145

Kurtosis	1.853433	1.220573

The results above show that the mean of the weekly total song units (2,558,209) is higher than that of the weekly total streams (1,702,839). The same also applies across their median (2,466,734 for weekly total song units, and 1,664,910 for weekly total streams), maximum (4,560,228 for weekly total song units, and 2,336,776 for weekly total streams) and minimum (1,634,875 for weekly total song units, and 1,412,225 for weekly total streams) values. Moreover, the range of values, that is the difference between the maximum and minimum values, for the weekly total song units (2,925,353) is larger than that of the weekly total streams (924,551), thus showing that the weekly total song unit distribution generally has a larger spread than that of the weekly total streams.

It is also evident that both the song unit and stream data are highly positively skewed (given the rule of thumb that highly skewed data typically exhibits a skewness whose value is absolutely greater than 1, whereas fairly symmetrical data typically exhibits a skewness whose value is absolutely less than 0.5). This means that for both data sets, their right-hand tail is longer than the left-hand tail, that is, the distribution of the data on each data set is shifted to the left, with its tail on the right side. This explains why the mean is greater than the median in both cases and further implies that there were both more song sales and streaming numbers in the last half of 2019 than the first half of 2020. It is presupposed that the drop in song sales and streaming numbers in 2020 seemed to have resulted from the impact of COVID-19. According to a report by Nielsen Music (2020), a lot of the key music listening hours, such as during commutes, were disrupted for many, thus explaining the reduced number of both streams and digital song units during that period. However, it is also evident that the distribution of stream data is more positively skewed than the distribution of the song unit data. This means that the distribution of stream data exhibited a longer right tail than that of the song unit data.

Additionally, both distributions exhibit a positive but small kurtosis value, meaning that both distributions exhibit fairly light tails. This means that fewer observations in both distributions are made at the tails, in comparison to a standard normal distribution, whose kurtosis value equals to three. In both cases, the distributions are thus referred to as platykurtic distributions as they have fewer outliers than an otherwise normal distribution.

4.2. Findings based on the Predictive Analysis.

4.2.1. Stationarity Tests

The periodical mean plots (based on the weekly total song unit and weekly total stream sample distributions, as shown on Appendix III & IV) propose a high likelihood that the data may be stationary in the long run. Nevertheless, both datasets are statistically tested for stationarity using the Augmented Dickey Fuller test to contrasting results, as summarized below.

Table 2: A table showing the Augmented Dickey Fuller test results for the Weekly Total Streams and Weekly Total Song unit distributions (The R outputs are shown on Appendix V).

	Weekly Total Song units	Weekly Total Streams
Augmented Dickey-Fuller	-1.668	-2.4701
Lag order	3	3
P-value	0.708	0.3853

The p-values for both tests (0.708 for the ADF test on the song unit time series, and 0.3853 for the ADF test on the stream time series) are greater than the significance level (0.05), revealing that the null hypothesis (the presence of a unit root) cannot be rejected. This implies that both time series are in fact not stationary. Both time series are hence transformed into temporal-independent time models and retested for stationarity.

Table 3: A table showing the Augmented Dickey Fuller test results for the Differenced Weekly Total Streams and Differenced Weekly Total Song unit distributions.

First Difference {I(1)}	Weekly Total Song units	Weekly Total Streams
Dickey-Fuller	-4.5107	-4.3617
Lag order	3	3
P-value	0.01	0.01

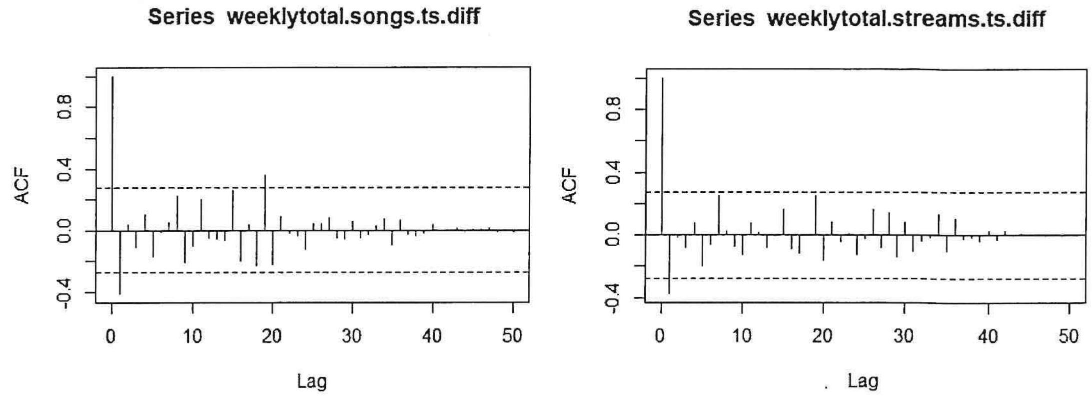
(The Dickey-Fuller test revealed that the actual p-values for both distributions were actually smaller than the printed values).

As shown above, the p-values for both tests are smaller than the significance level (0.05), revealing that the null hypothesis (the presence of a unit root) is thus rejected. This implies that the differenced time series may be stationary or trend stationary.

4.2.2. Model Type and Lag Order Identification

To determine the lag order, each independent time series is subjected to Partial Autocorrelation (PACF) and Autocorrelation (ACF) functions. The partial autocorrelation functions for both the song unit and stream time series (as shown in Appendix VII and VIII) seem to geometrically decrease towards 0 overtime, thus indicating that both models are Moving Average models.

Figure 7: Two graphs showing the ACF functions for both the Differenced Weekly Total Song units and the Differenced Weekly Total Streams (The equivalent PACFs are shown in Appendix VII & VIII)



For the song unit time series, the sample ACF curve reveals a statistically significant spike at lag 1, followed by non-significant values for other lags. In this case, the spike taken at lag 19 is regarded as a sampling error, thus considered non-significant. This indicates that the model is an MA (1) process.

For the stream time series, the sample ACF curve reveals a statistically significant spike at lag 1, followed by non-significant values for other lags. This indicates that this model is also an MA (1) process. Both models thus have a (p,d,q) combination of (0,1,1) as they both require 1 difference to make the series stationary and are a Moving average model of lag (1). This means that the Moving Average models for both time series can be formalized as:

$$S_t' = \beta_0 + \varepsilon_t + \beta_1\varepsilon_{t-1}$$

for the stream time series, where St represents the current stream-equivalent units, ε_t represents the current stream error, ε_{t-1} represents the lag (1) stream error and where ε_t is assumed to be identically and independently distributed with a normal distribution; and:

$$D_t' = \alpha_0 + \theta_t + \alpha_1\theta_{t-1}$$

for the song unit time series, where Dt represents the current digital song units, θ_t represents the current song unit error, θ_{t-1} represents the lag (1) song unit error and where θ_t is also assumed to be identically and independently distributed with a normal distribution.

4.2.3. The Akaike Information Criteria

Table 4: A table showing the Akaike Information Criteria results for both the various Differenced Weekly Total Song units and the Differenced Weekly Total Streams of orders I(1) and I(2).

Orders	Differenced Weekly Total Song units	Differenced Weekly Total Streams
I(1)	1438.838	1391.98
I(2)	1487.345	1394.144

The Akaike Information Criteria (AIC) further reveals that both the song unit and stream time series possess an optimal lag length of order (1), as shown in the table above. This is because in both cases, the AIC measures of the independent ARIMA (0,1,1) models are significantly lower than that of their relative ARIMA (0,1,2) models, making the ARIMA (0,1,1) models the more suitable models to use.

4.2.4. Coefficient Estimation

Using the ARIMA function in R, the coefficients (β and α) for the stream and song unit time series can respectively be estimated. The model specification for the differenced song unit time series are thus estimated as follows.

$$D_t' = -5766.36 + \theta_t - 0.6153\theta_{t-1}$$

The above equation thus shows that an inverse relationship exists between the change in number of digital song unit sales and the previous period's lag error, such that a unit increase in the last period's lag error may cause a decrease in the difference between the number of song unit sales in the previous period (D_{t-1}) and the current number of song units (D_t), by a factor equivalent to 0.6153. Moreover, the negative constant (-5766.36) implies a general tendency for future numbers of song unit sales to be less than their equivalent previous numbers – that is, D_t tends to be less than D_{t-1} for subsequent periods. This holds true unless the value of the error term, θ_t for the subsequent period(s), is more than the constant for that period. However, the latter assumes that the lagged period's error is equal to 0.

The model specification for the differenced stream time series can thus be estimated as follows.

$$S_t = -4492.394 + \varepsilon_t - 0.6110\varepsilon_{t-1}$$

As with the song unit time series, the above equation thus shows that an inverse relationship exists between the change in number of streams and the previous period's lag error, such that an increase in the last period's lag error may cause a decrease in the difference between the number of streams in the previous period (S_{t-1}) and the current number of streams (S_t), by a factor whose value is equivalent to 0.6110. Moreover, the negative constant (-4492.394) implies a general tendency for future numbers of streams to be less than their equivalent previous numbers – that is, S_t tends to be less than S_{t-1} for subsequent periods. This holds true unless the value of error term, ε_t for the subsequent period(s), is more than the constant for that period. However, the latter assumes that the lagged period's error is equal to 0.

4.2.5. The Ljung-Box Test

Table 4: A table showing Ljung Box test results for both the Differenced Weekly Total Song units and the Differenced Weekly Total Streams.

Orders	Differenced Weekly Total Song units	Differenced Weekly Total Streams
X-squared	0.13007	0.45542
P-value	0.7184	0.4998

The p-values for both tests (0.7184 for the Ljung-Box test on the song unit time series, and 0.4998 for the Ljung-Box test on the stream time series) are greater than the significance level (0.05), revealing that the null hypothesis – the model does not show lack of fit – cannot be rejected. This means that not only do the models have statistically sufficient lag orders, but also exhibit white noise residuals – that is, no serial correlation exists – thus making the existing coefficients also efficient.

4.2.6. Forecasting

The models above are hence used to predict future values of individual data with a 95% confidence level, as shown below.

Figure 8: A graph showing the Forecasts for Future Song units with non-zero mean.

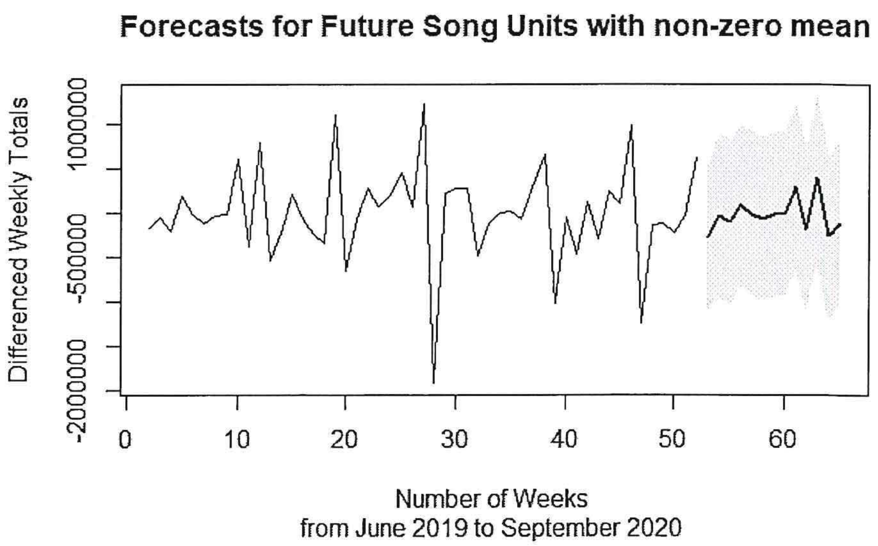
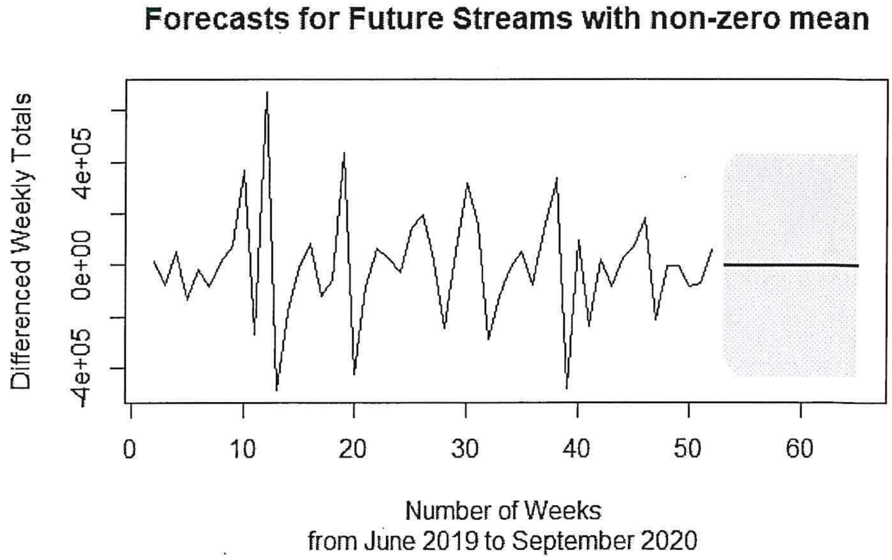


Figure 9: A graph showing the Forecasts for Future Streams with non-zero mean.

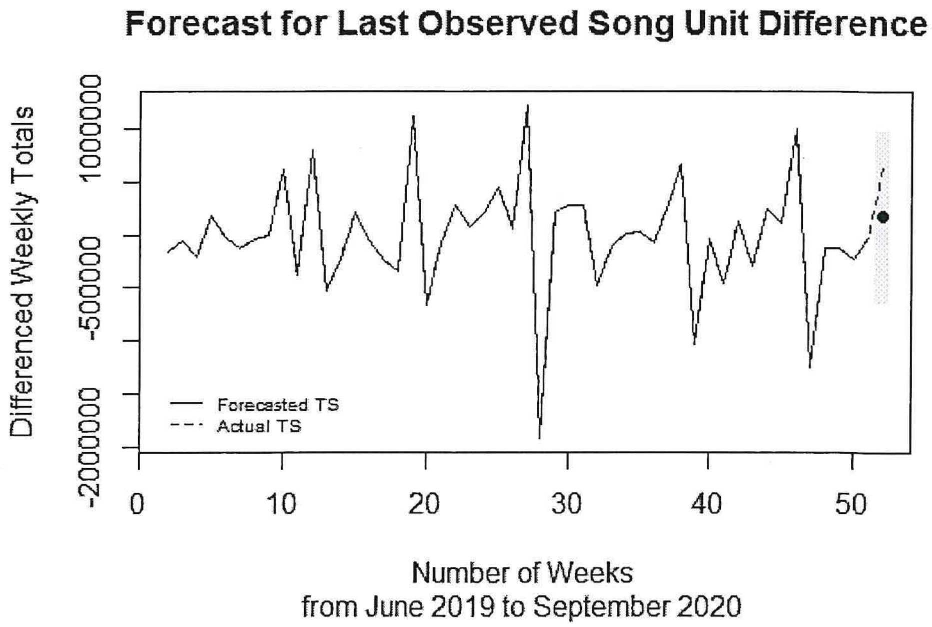


From the above graphs, we can observe that song unit differences may be forecasted overtime with variable future values being returned over the additional 13-week window. However, the stream difference time series model returns a flat forecast. On one hand, this could mean that future streaming observations may exhibit random walk properties instead of the normal properties as initially assumed for this study. On the other hand, it could simply mean that the sample data used for this data contains insufficient temporal characteristics (such as no definitive trend or seasonality) to allow the model to be able to derive future observations that have different conditional means.

4.2.7. Forecasting Accuracy

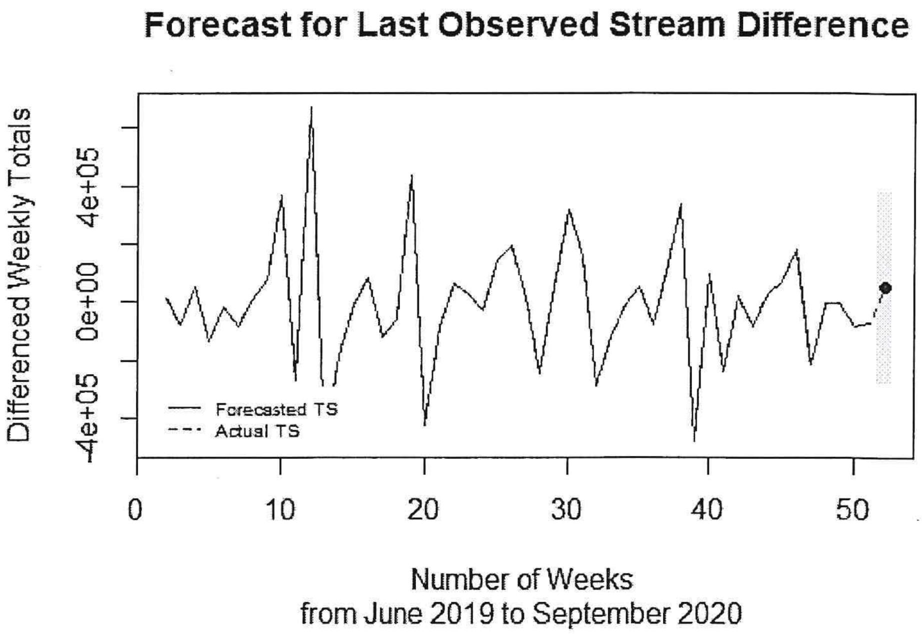
To examine the forecasting accuracy, the last known week's song unit and stream difference is removed from each dataset to form a new dataset. It is this new dataset that uses the respective times series models to forecast the data for the last known period.

Figure 10: A graph showing the Forecasts for the Last Observed Song Unit Difference.



For the song unit differences, the actual observed song unit difference from the sample set used is 622,664.6 whereas the difference predicted by the equivalent moving average model comes to 172,780.1, resulting in an estimation error of 449,884.5. This disparity seems to indicate that the time series model, based on the song unit sample, is fairly inaccurate when it comes to predicting near future values.

Figure 11: A graph showing the Forecasts for the Last Observed Stream Difference.



For the stream differences however, the actual observed song unit difference from the sample set used is 60,335.39 whereas the difference predicted by the equivalent moving average model comes to 49,024.93, resulting in an estimation error of 11,310.46. This difference, in comparison to that of the song unit difference, seems to indicate that the time series model, based on the stream sample, is more accurate than the song unit time series model when it comes to predicting near future values. However, as aforementioned, the flat forecast exhibited by the stream difference time series model seems to allude to random walk properties instead of the normal properties as initially assumed for this study, thus proposing the likelihood that both models might in actuality be random walk series in the short term.

4.3. Findings based on the Correlational Analysis.

4.3.1. The VAR Model

Carrying forward the properties of the previous two MA (1) models, the following bi-directional vector autoregressive model is formulated:

$$S_t' = \beta_{10} + \beta_{11}\varepsilon_{1t-1} + \beta_{12}\varepsilon_{2t-1} + \varepsilon_{1t}$$

$$D_t' = \beta_{20} + \beta_{21}\varepsilon_{2t-1} + \beta_{22}\varepsilon_{1t-1} + \varepsilon_{2t}$$

where D_t represents the current digital song units, S_t represents the current stream-equivalent units, ε_{1t} represents the current stream error, ε_{2t} represents the current song unit error, ε_{1t-1} represents the lag (1) stream error, ε_{2t-1} represents the lag (1) song unit error, and where both ε_{1t} and ε_{2t} are assumed to be identically and independently distributed with a normal distribution.

4.3.2. Coefficient Estimation

Using the VAR function in R, the coefficients (β_{it}) for the stream and song unit time series can respectively be estimated. The model specification for the vector autoregressive model can thus be estimated as follows.

$$S_t' = -8023.859 - 0.2786821\varepsilon_{1t-1} - 0.05133454\varepsilon_{2t-1} + \varepsilon_{1t}$$

$$D_t' = -6835.5414988 - 0.5892953\varepsilon_{2t-1} + 0.5666045\varepsilon_{1t-1} + \varepsilon_{2t}$$

The first half of the equation looks into the effect of the song unit lag error and the streaming lag error on the change of streaming numbers. The equation thus shows that an inverse relationship seems to exist between the change in number of streams and the previous period's song unit and streaming lag errors. This is such that a unit increase in the previous period's song unit lag error may cause a decrease in the difference between the number of streams in the previous period (S_{t-1}) and the current number of streams (S_t), by a factor equivalent to 0.05133454, whereas a unit increase in the previous period's streaming lag error may cause a decrease in the difference between the number of streams in the previous period and the current number of streams, by a factor equivalent to 0.2786821. From this, it is evident that the streaming lag error has a greater effect on the change of streaming numbers than the song unit lag error, as its coefficient (-0.2786821) is larger than that of the song unit lag error (-0.05133454). As with the equivalent independent time series, the negative constant (-8023.859) implies a general tendency for future numbers of streams to be less than their previous numbers – that is, S_t tends to be less than S_{t-1} for subsequent periods. This holds true unless the value of the error term, ε_t for the subsequent period(s), is more than the constant for that period. However, the latter assumes that the lagged period's song unit and streaming errors both equal to 0.

The second half of the equation looks into the effect of the song unit lag error and the streaming lag error on the change of digital song unit sales. The equation thus shows that an inverse relationship seems to exist between the change in number of song units and the previous period's song unit error, such that a unit increase in the previous period's song unit lag error may cause a decrease in the difference between the number of song units in the previous period (D_{t-1}) and the current number of song units (D_t), by a factor equivalent to 0.5892953. However, the equation also shows that a direct relationship seems to exist between the change in number of song units and the previous period's streaming lag error, such that a unit increase in the previous period's streaming lag error may cause an increase in the difference between the number of song units in the previous period and the current number of song units, by a factor equivalent to 0.5666045. Nevertheless, it is evident that the song unit lag error has a greater effect on the change of streaming numbers than the song unit lag error, as its coefficient (-0.5892953) is larger than that of the song unit lag error (+0.5666045). As with the equivalent independent time series, the negative constant (-6835.5414988) implies a general tendency for future numbers of streams to be less than their previous numbers – that is, D_t tends to be less than D_{t-1} for subsequent periods. This holds true unless the value of error terms, ε_t for the subsequent period(s) and ε_{1t-1} (streaming error) for the subsequent first lag periods, are both more than the constant for that period. However, the latter assumes that the lagged period's song unit lag error is equal to 0.

4.3.3. The Granger-Causality Test

Table 4: A table showing the Granger-Causality test results for both the various Differenced Weekly Total Song units and the Differenced Weekly Total Streams.

Cause Variable =	**Differenced Weekly Total Songs**	**Differenced Weekly Total Streams**

F-test	0.39951	1.3165
P-value	0.5289	0.2541

The Granger-Causality test, in this case, simultaneously investigates whether the Song unit differences Granger-cause (have an effect on) the Streaming differences and whether the Streaming differences Granger-cause the Song unit differences. Using the causality function in R, the test revealed that the p-values for both causations (0.5289 and 0.2541) are more than the significance level (0.05). This means that the null hypothesis in both cases cannot be rejected and thus, all the coefficients from the vector autoregressive model are equal to zero. This further implies that streaming does not Granger-cause digital song unit sales as its past values do not contain enough information to predict current or future song units. This also implies that song units do not Granger-cause streaming numbers as its past values do not contain enough information to predict current or future streaming numbers. Both time series are thus unrelated and independent, and thus revokes any kind of relationship between the two variables.

CHAPTER 5: DISCUSSION AND CONCLUSION

5.1. Discussion

Based on the weekly U.S. chart data of the Top 200 most streamed songs from Spotify and weekly U.S. rank data of the Top 100 most sold songs, as depicted by the Rolling Stones Music Charts, an analysis of the two independent univariate time series models reveals that the future values of both streaming and song unit variables cannot accurately be forecasted. This is further emphasized by the flat forecast depicted on the first difference univariate moving average model that is developed for the streaming distribution. As aforementioned, the flat forecast suggests that the streaming distribution may possibly be following a random walk process (in the short run), as opposed to the time series process as initially established. Furthermore, the disparity that comes from the results of the accuracy analysis of the song unit difference time series model seems to propel this conclusion, thus indicating the possibility that the song unit differences might also, in actuality, follow a random walk process.

The plausible impact of streaming services on record sales is also investigated by subjecting a Granger-Causality test on a vector autoregressive model. Given the p-values for both causations (that is, the streaming differences Granger-causing the song unit differences and vice versa) were both more than the significance level (0.05), the null hypothesis is thus rejected in both cases, hence revealing that there is no causation element between the two variables. This implies the distribution processes for both the streaming numbers and digital song unit sales are not only seemingly unpredictable but also independent of each other, thus highlighting that there is no relationship between the two variables. Therefore, the results show that neither a substitution nor complementary effect seems to exist between streaming and digital record sales, and future observations of both variables seem to be random.

5.2. Implications and Contributions

The results demonstrated in this project seem to provide several implications for research. Predominantly, the findings go against the overall view that streaming services substitute or complement record sales, and instead reduces streaming to a distribution medium rather than a determinant of music revenue. Secondly, different from previous studies that examined the relationship between various streaming services and physical music consumption or streaming services and physical record sales, this study looks into the relationship between streaming services and purely digital music sales. Thirdly, different from other more generalized studies, this project uses a case study of Spotify to conduct a more in-depth analysis of the relationship between standalone streaming services on digital record sales, with Spotify as the more suitable proxy for standalone streaming services. Furthermore, contrary to previous research, this study uses actual streaming numbers and approximated digital song sales as the more suitable proxy for digital music industry revenue, instead of general music consumption, thus providing more dependable results. Lastly, this project takes a look at the relationship between standalone streaming services and digital record sales in the United States. According to the e International Federation of the Phonographic Industry (2019), the United States of America is considered the largest music market in the world, and thus is the country that has the greatest contribution to the impact of music globally.

5.3. Limitations, Recommendations and Future Research Opportunities

Despite the above contributions, this study also contains several limitations. Predominantly, the sample period used for this study only allowed for short term effects to be observed, thus limiting the conciseness of the conclusions made with regard to long term effects. Secondly, the conclusions made in this project do not consider the artists-specific characteristics that may affect the output of data. At the beginning of this study, it is assumed that all the data output comes from one record label. This is a major consumption considering that it is the norm for artists and songs to be released and managed under different record labels. Furthermore, using the weekly totals of both the streaming numbers and song unit sales only generalizes the research topic, thus limiting the influence of specific artist contributions to the sample. Future research may also incorporate the effects of different record labels as well as other possible determinants such as the audience demographic, gender of the artist, public reception of the songs, amongst other artist-specific elements to better explain the behaviour of the variables over time.

In conclusion, this project proves that there is no time-dependent relationship between standalone streaming numbers and record sales. Furthermore, it proves that it is difficult to accurately predict both streaming numbers and song sales based on short-term information. From this, it is recommended that future research could instead try to focus on the effects of the two based on long-term data. It is also recommended that further research could try and prove if both song sale units and streaming numbers do in fact follow a random walk process.

REFERENCES

Wilkinson, E. (2000). *Chinese History: A Manual, 2nd edition, Revised and Enlarged.* Cambridge, MA: Harvard University Asia Centre distributed by Harvard University Press. ISBN 0674002490.

Orpheus, R. (2017). *The Economics of the Music Industry and the Impact of Digital Technology*. Medium. Retrieved from https://medium.com/@rodneyorpheus/the-digital-music-industry-an-overview-965915ae1fab

Recording Industry Association of America. (2019). *Year End 2019 RIAA Music Revenues Report*. Retrieved from https://www.riaa.com/wp-content/uploads/2020/02/RIAA-2019-Year-End-Music-Industry-Revenue-Report.pdf

Gamal, A. E. (2012). *The Evolution of the Music Industry in the Post-Internet Era*. Retrieved from https://scholarship.claremont.edu/cgi/viewcontent.cgi?article=1501&context=cmc_theses

Peitz, M., and Waelbroeck, P. (2004). *The Effect of Internet Piracy on CD Sales: Cross-Section Evidence*. Retrieved from https://www.researchgate.net/publication/5001629

Lee, M., Choi, H., Cho, D., and Lee, H. (2016). *Cannibalizing or Complementing? The Impact of Online Streaming Services on Music Record Sales*. Retrieved from https://reader.elsevier.com/reader/sd/pii/S187705091631359X?token=1297901AB5C2155B6870BD26A81EA783EA2C485A6CE361014F614C5550917C304F4E3C8A5E1595D737A755FA84318C9E

Desai, M. (2019). *How Spotify makes Money – Business Model.* Medium. Retrieved from https://medium.com/dissecting-music-tech/how-spotify-makes-money-business-model-ca0a71a19163

Goldman, D. (2010). *Music's lost decade: Sales cut in half.* CNN Money. Retrieved from https://money.cnn.com/2010/02/02/news/companies/napster_music_industry/

Michel, N. J. (2006). *The Impact of Digital File Sharing on the Music Industry: An Empirical Analysis*. Topics in Economic Analysis & Policy, 6(1). Retrieved from https://www.riaa.com/wp-content/uploads/2004/01/ art-the-impact-of-digital-file-sharing-on-the-music-industry-michel-2006.pdf

Liebowitz, S. J. (2005). *Economists Examine File-Sharing and Music Sales.* Retrieved from https://www.researchgate.net/publication/23745956

Smith, M. D., and Telang, R. (2009). *Competing With Free: The Impact of Movie Broadcasts on DVD Sales and Internet Piracy*. SSRN Electronic Journal, 33(2), 321-338. Retrieved from https://www.researchgate.net/publication/220260253

Nguyen, G. D., Dejean, S., and Moreau, F. (2013). *On the complementarity between online and offline music consumption: the case of free streaming*. Journal of Cultural Economics, 38(4). Retrieved from http://www.researchgate.net/publication/257578428

Aguiar, L. (2017). *Let the music play? Free streaming and its effects on digital music consumption*. Retrieved from https://reader.elsevier.com/reader/sd/pii/S016762451630110X ?token=BECF03936C07D285C1314F2EB6DB3559C9BA77A48A0BDB96AEDAE861E77 CE1339AE193DA4FBB92379D483168F943BAF6

Aguiar, L., and Waldfogel, J. (2015). *Streaming Reaches Flood Stage: Does Spotify Stimulate or Depress Music Sales?*. Retrieved from https://www.nber.org/papers/w21653.pdf

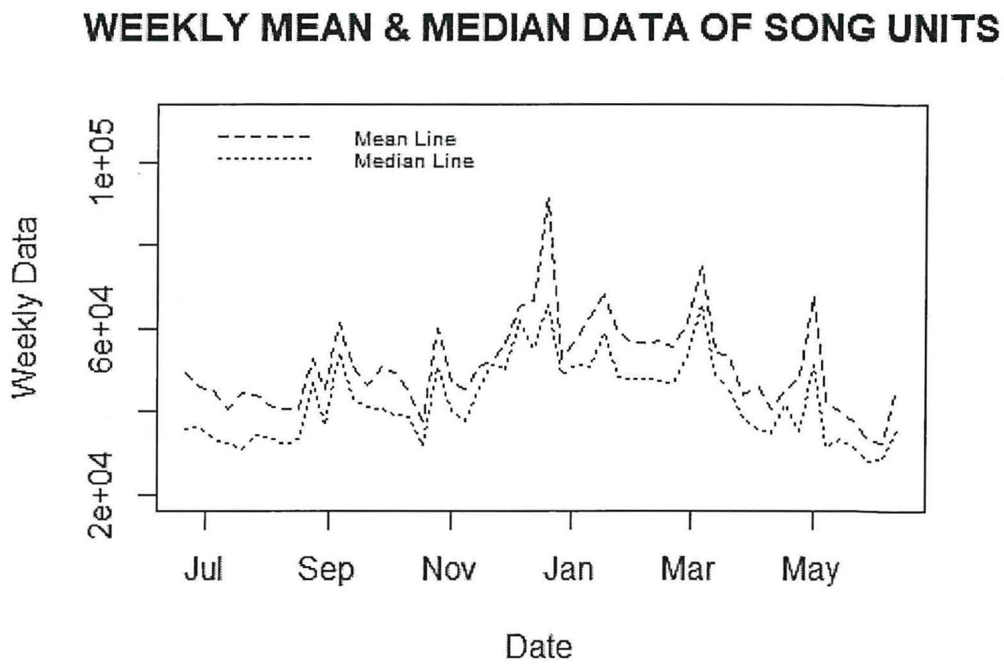
Bukszpan, D. (2019). *Vinyl records are ontrack to outsell compact discs for the first time in 33 years and prices have risen 490%*. CNBC. Retrieved from https://www.cnbc.com /2019/09/14/vinyl-records-on-track-to-outsell-cds-and-prices-have-risen-490percent.html

Field, M. (2020). *The best music streaming services: Apple Music, Spotify, YouTube Music and Amazon Music compared*. The Telegraph. Retrieved from https://www.telegraph.co.uk/ technology/0/best-music-streaming-services-apple-music-spotify-amazon-music/#:~:text=Spotify%20is%20the%20largest%20subscription,you%20get%20with%20Apple%20Music.

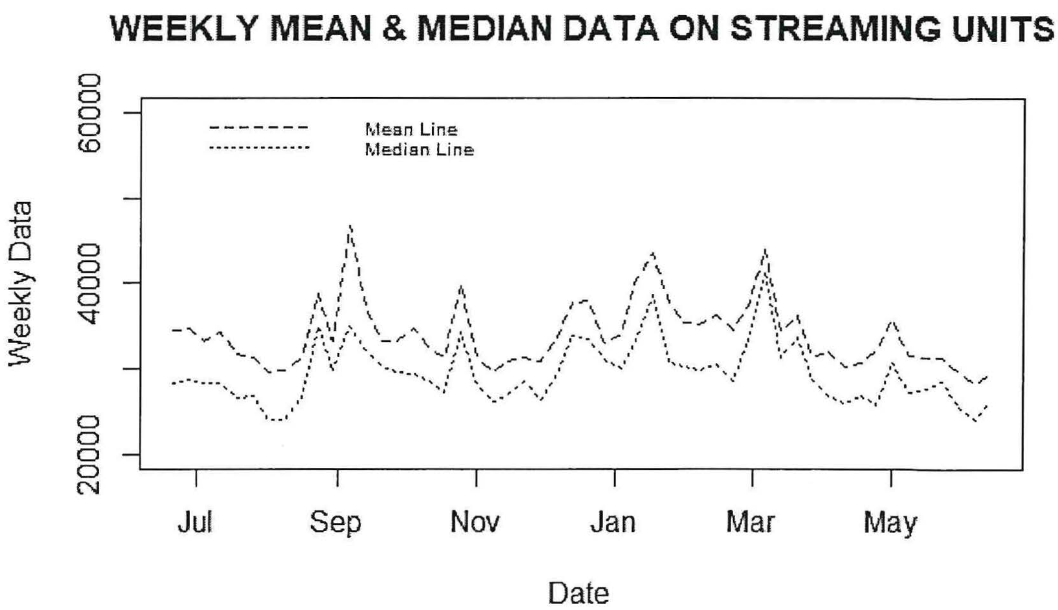
Recording Industry Association of America. (2016). *Digital Single Award RIAA and GR&F Certification Audit Requirements.* Retrieved from https://www.riaa.com/wp-content/uploads/2016/02/DIGITAL-SINGLE- AWARD-RIAA-AND-GRF-CERTIFICATION-AUDIT-REQUIREMENTS.pdf

International Federation of the Phonographic Industry. (2019). *Global Music Report: The Industry in 2019.* Retrieved from https://www.ifpi.org/wp-content/uploads/2020/07/Global_Music_Report-the_Industry_in_2019-en.pdf

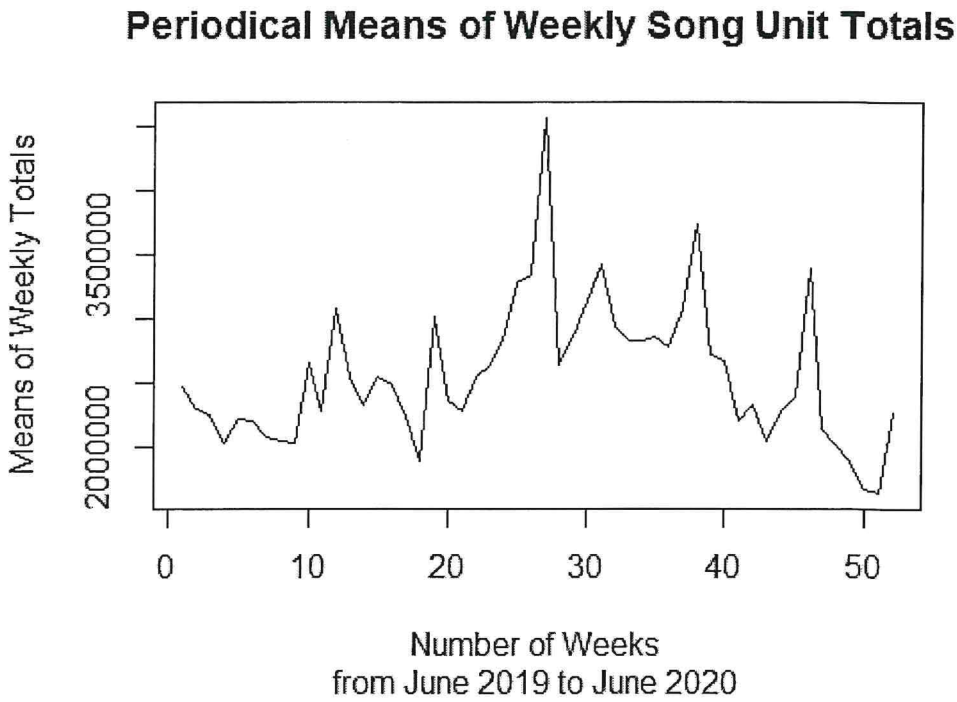
APPENDICES



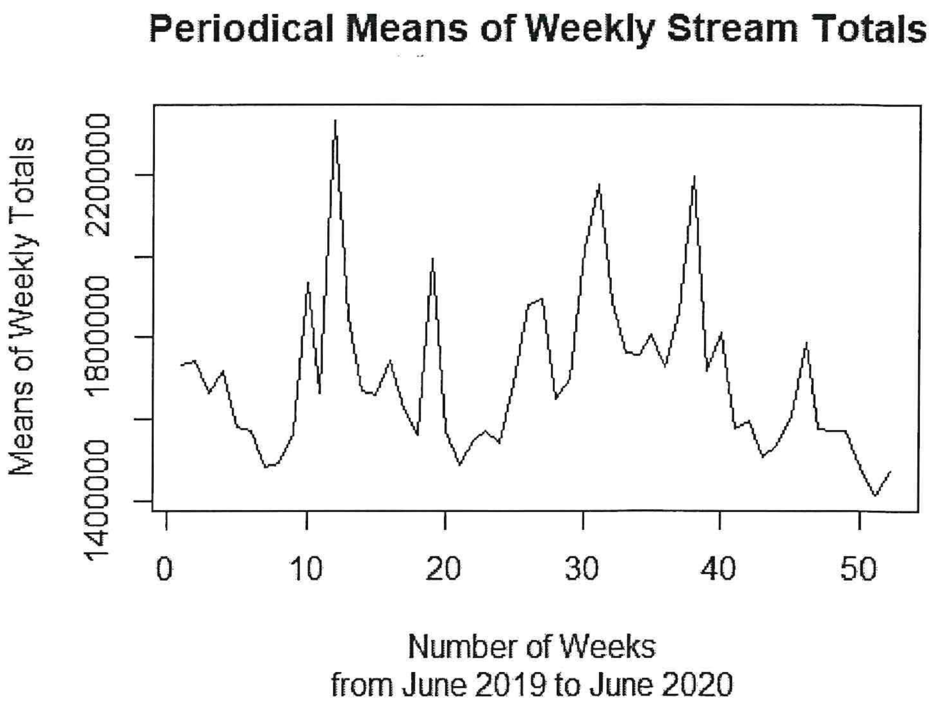
Appendix I: The Weekly Means & Medians of Song Units



Appendix II: The Weekly Means & Medians of Streams



Appendix III: The Periodical Means of Weekly Song unit Totals



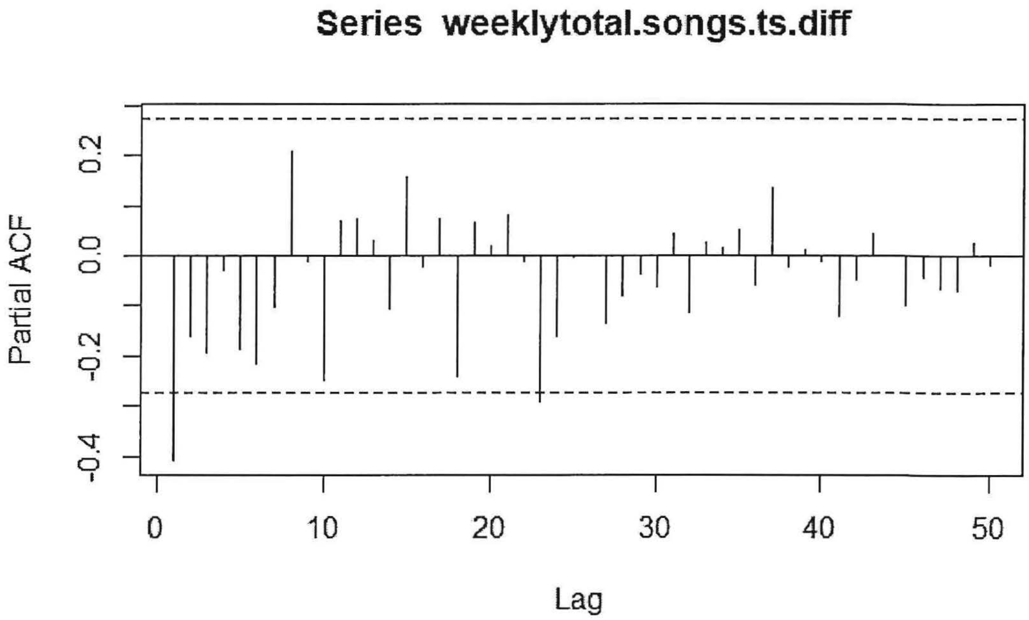
Appendix IV: The Periodical Means of Weekly Stream Totals

```
> adf.test(weeklytotal.songs.ts, alternative = c("stationary", "explosive"), k = trunc((length(weeklytotal.songs.ts)-1)^
(1/3)))

        Augmented Dickey-Fuller Test

data:  weeklytotal.songs.ts
Dickey-Fuller = -1.668, Lag order = 3, p-value = 0.708
alternative hypothesis: stationary

> adf.test(weeklytotal.streams.ts, alternative = c("stationary", "explosive"), k = trunc((length(weeklytotal.streams.ts)-
1)^(1/3)))

        Augmented Dickey-Fuller Test

data:  weeklytotal.streams.ts
Dickey-Fuller = -2.4701, Lag order = 3, p-value = 0.3853
alternative hypothesis: stationary
```

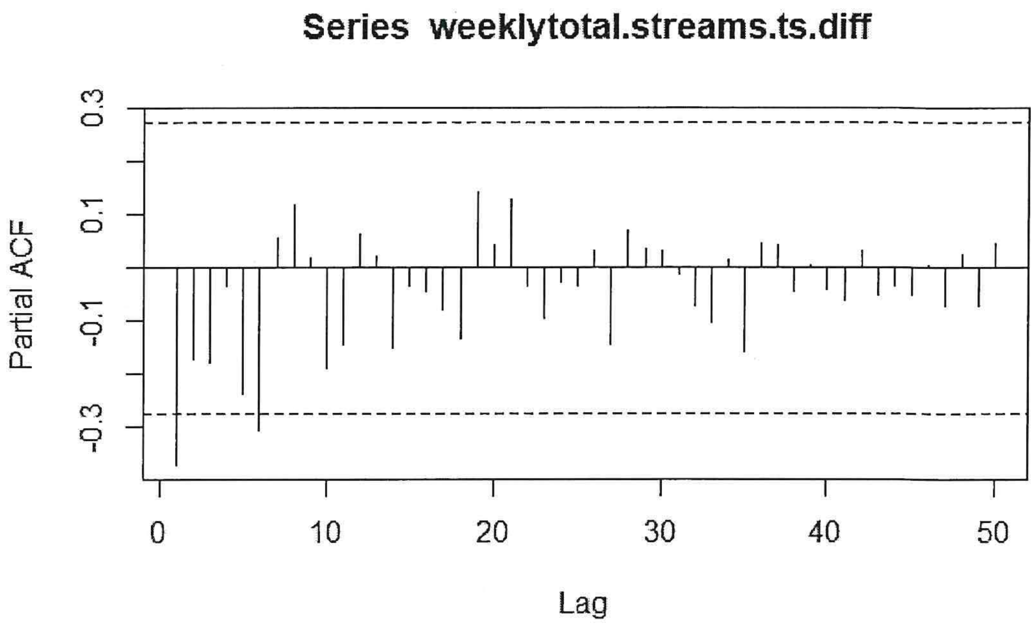
Appendix V: The Results from the Dickey-Fuller Tests for both the Weekly Total Song unit and Weekly Total Stream Distributions, as run on R.

```
> adf.test(weeklytotal.songs.ts.diff, alternative = c("stationary", "explosive"), k = trunc((length(weeklytotal.songs.ts.d
iff)-1)^(1/3)))

        Augmented Dickey-Fuller Test

data:  weeklytotal.songs.ts.diff
Dickey-Fuller = -4.5107, Lag order = 3, p-value = 0.01
alternative hypothesis: stationary

Warning message:
In adf.test(weeklytotal.songs.ts.diff, alternative = c("stationary",  :
  p-value smaller than printed p-value
> adf.test(weeklytotal.streams.ts.diff, alternative = c("stationary", "explosive"), k = trunc((length(weeklytotal.streams.
ts.diff)-1)^(1/3)))

        Augmented Dickey-Fuller Test

data:  weeklytotal.streams.ts.diff
Dickey-Fuller = -4.3617, Lag order = 3, p-value = 0.01
alternative hypothesis: stationary

Warning message:
In adf.test(weeklytotal.streams.ts.diff, alternative = c("stationary",  :
  p-value smaller than printed p-value
```

Appendix VI: The Results from the Dickey-Fuller Tests for both the Differenced Weekly Total Song unit and Differenced Weekly Total Stream Distributions, as run on R.



Appendix VII: Partial Autocorrelation function plot for the Differenced Weekly Song Totals



Appendix VIII: Partial Autocorrelation function plot for the Differenced Weekly Stream Totals

```
> arima_songs <- arima(weeklytotal.songs.ts.diff, order = c(0,0,1), seasonal = list(order
 = c(0,0,1), period = 51), include.mean = TRUE)
> arima_songs2 <- arima(weeklytotal.songs.ts.diff, order = c(0,0,2), seasonal = list(orde
r = c(0,0,2), period = 51), include.mean = TRUE)
> AIC(arima_songs)
[1] 1483.838
> AIC(arima_songs2)
[1] 1487.345
> arima_streams <- arima(weeklytotal.streams.ts.diff, order = c(0,0,1), seasonal = list(o
rder = c(0,0,1), period = 51), include.mean = TRUE)
> arima_streams2 <- arima(weeklytotal.streams.ts.diff, order = c(0,0,2), seasonal = list
(order = c(0,0,2), period = 51), include.mean = TRUE)
> AIC(arima_streams)
[1] 1391.98
> AIC(arima_streams2)
[1] 1394.144
```

Appendix IX: The Akaike Information Criterion on both the Song unit and Stream Time Series

```
> arima_songs

Call:
arima(x = weeklytotal.songs.ts.diff, order = c(0, 0, 1), seasonal = list(order = c(0,
    0, 1), period = 51), include.mean = TRUE)

Coefficients:
          ma1    sma1  intercept
      -0.6153  0.9372   -5766.36
s.e.   0.1274  2.2776   25552.45

sigma^2 estimated as 1.143e+11:  log likelihood = -737.92,  aic = 1483.84
```

Appendix X: The ARIMA (0, 1, 1) model based on the Weekly Total Song units' sample distribution.

```
> arima_streams

Call:
arima(x = weeklytotal.streams.ts.diff, order = c(0, 0, 1), seasonal = list(order = c(0,
    0, 1), period = 51), include.mean = TRUE)

Coefficients:
          ma1     sma1  intercept
      -0.6110  -7e-04  -4492.394
s.e.   0.1659     NaN  10601.458

sigma^2 estimated as 3.539e+10:  log likelihood = -691.99,  aic = 1391.98
```

Appendix XI: The ARIMA (0, 1, 1) model based on the Weekly Total Streams' sample distribution.

```
> Box.test(resid(arima_songs),type = "Ljung", lag = 1)

        Box-Ljung test

data:  resid(arima_songs)
X-squared = 0.13007, df = 1, p-value = 0.7184

> Box.test(resid(arima_streams),type = "Ljung", lag = 1)

        Box-Ljung test

data:  resid(arima_streams)
X-squared = 0.45542, df = 1, p-value = 0.4998
```

Appendix XII: The Ljung-Box Tests on both the Song unit and Stream Time Series

```
VAR Estimation Results:
=======================

Estimated coefficients for equation weeklytotal.songs.ts.diff:
==============================================================
Call:
weeklytotal.songs.ts.diff = weeklytotal.songs.ts.diff.l1 + weeklytotal.streams.ts.diff.l1 + const

  weeklytotal.songs.ts.diff.l1 weeklytotal.streams.ts.diff.l1                          const
                    -0.5892953                      0.5666045                  -6835.5414988

Estimated coefficients for equation weeklytotal.streams.ts.diff:
================================================================
Call:
weeklytotal.streams.ts.diff = weeklytotal.songs.ts.diff.l1 + weeklytotal.streams.ts.diff.l1 + const

  weeklytotal.songs.ts.diff.l1 weeklytotal.streams.ts.diff.l1                          const
                 -5.133454e-02                  -2.786821e-01                  -8.023859e+03
```

Appendix XIII: The regression results for the Vector Autoregression Model

```
> vars::causality(Combined_VAR, cause = "weeklytotal.songs.ts.diff")
$Granger

	Granger causality H0: weeklytotal.songs.ts.diff do not Granger-cause
	weeklytotal.streams.ts.diff

data:  VAR object Combined_VAR
F-Test = 0.39951, df1 = 1, df2 = 94, p-value = 0.5289


$Instant

	H0: No instantaneous causality between: weeklytotal.songs.ts.diff and
	weeklytotal.streams.ts.diff

data:  VAR object Combined_VAR
Chi-squared = 18.33, df = 1, p-value = 1.858e-05


> vars::causality(Combined_VAR, cause = "weeklytotal.streams.ts.diff")
$Granger

	Granger causality H0: weeklytotal.streams.ts.diff do not Granger-cause
	weeklytotal.songs.ts.diff

data:  VAR object Combined_VAR
F-Test = 1.3165, df1 = 1, df2 = 94, p-value = 0.2541


$Instant

	H0: No instantaneous causality between: weeklytotal.streams.ts.diff and
	weeklytotal.songs.ts.diff

data:  VAR object Combined_VAR
Chi-squared = 18.33, df = 1, p-value = 1.858e-05
```

Appendix XIV: The Granger Causality Tests on the VAR Model