National College of Ireland

BSc(Hons) in Computing

Data Analytics

2020/2021

Carl Flynn

X17347726

X17347726@student.ncirl.ie

An Analysis of the Shift in Music Industry Revenue Streams in the Last 15 Years

Technical Report

Contents

Executive Summary

This report details how this research project was undertaken, the methodology used and an analysis of how music consumption trends have changed in recent years and how this has subsequently affected music industry revenue. The overall aim of this project is to gather and analyse appropriate data on the music industry from recent years and draw conclusions from this data using various analysis techniques as well as forecast into the future to see what kind of position the music industry will be in the next three years. This report details the data used in the research, the methodology used and how it was implemented and all results found. Various datasets were gathered online from websites such as Kaggle.com and statista.com covering a range of subjects like revenue, format market shares and sales and also covering several nations/regions as well as individual companies within the music industry. This analysis found that over the selected time scale there has been a universal shift towards digital music consumption formats and that music industry revenue across regions and companies and industry as a whole, has been on the rise in recent years. Using forecasting techniques we have also made predicted to continue rising over the next three years based on previous years and it was concluded that music consumption will likely continue to be dominated by digital mediums in the coming years.

1.0   Introduction

1.1. Background

While looking for topics to choose for this project I was stuck looking online at random datasets and nothing of interest seemed to be jumping out for me, after this I decided to rethink how I was searching for a topic and began looking at my own interests and hobbies for inspiration. Music seemed like the obvious choice as it has always been something I turn to whether I'm happy, sad, bored, busy, walking or relaxing, music has always been a mainstay. Since beginning college, I took a keen interest in DJing and have learned and progressed a great deal these last few years while also learning more about musical theory and structure. Being a computing student this also inevitable led me down the path of looking at the technical side of music like how MP3 files differ from WAV files, why vinyl sound quality still holds up in the midst of technological advances, how streaming services use the data gathered from users and how to view the analytics and the list goes on.

Being born in the late 90's I also remember first-hand going through many ways of listening to music over the years, cassette tapes and CD's piled high in a random cupboard in the house with a few still lingering around even today, then getting my first iPod and buying songs from the iTunes store, never doing anything as dangerous as downloading MP3's from a dodgy website online and finally where I am today using streaming services such as Spotify and Soundcloud as my main forms of listening to music, all while paying €5 (Thank you student subscription) for no ads while listening.

While racking my brains for a specific topic within music this seemed of interest while also feeling personal. Upon further research a number of useful resources and available

datasets appeared which provided me with sufficient reason to undertake this project. As someone who loves music I wanted to see where I fit in with the change in consumption over the years as well as learn about the music industry as a whole in recent years, this project will be of interest to anyone who enjoys music and is interested in some of the consumer insights and maybe even people beyond this.

1.2. Aims

1. To find adequate datasets that are relevant to the area in the music industry I am covering. Revenue, music format and sales data are all key areas that are essential to the project. Websites like Kaggle, Statista and RIAA are all reliable sources.
2. To clean and extract the needed data from the gathered data. This includes removing rows/columns that may contain years outside of the range covered in this project, removing unnecessary sheets from data gathered from Statista and removing any invalid entries that may be present in any of the datasets
3. To combine similar datasets into larger compiled datasets so the data is easier to examine. This reduces the total number of datasets considerably and makes it easier to perform tests and analysis.
4. To begin analysing the data and fulfilling the objectives of the project. This includes generating visualisations and performing tests on the data to meet our objectives.
5. To record results and examine as necessary to draw conclusions and make sure the analysis was successful.
6. To complete all documentation and arrange all results and visualisations accordingly so the reader can easily follow.

1.3. Technology

R Studio: R Studio is the main piece of software used for combining, cleaning, manipulating, and storing the gathered data. R Studio is an open-source integrated development environment for the R programming language which is the main programming language used in this project. R Studio contains a number of packages which provide extra functions for the aforementioned tasks as well as analysis statistical tests and data visualisation. The main package used in this package was the Tidyverse package which is a parent package containing a number of useful packages like ggplot2, dplyr and tidyr which all have their individual uses.

R Language: This is the main language used in the project with all steps of the KDD methodology using R in some aspect.

Excel: Spreadsheet tool with built in statistical and data visualisation functions. Mainly used for viewing data in spreadsheet format, performing some file manipulating/cleaning, running some statistical tests and generating quick and easy to read graphs to confirm findings from other tools.

SPSS: Statistical analysis software used for running statistical tests, forecasting and data visualisations. SPSS was used for generating clear, concise, and polished graphs and charts as well as performing relevant statistical tests with relative ease and returning user friendly, easy to read results and visualisations.

PowerBI: A collection of software services and apps provided by Microsoft used to transform data into interactive reports, in this project it was primarily used for generating visualisations for analysis and use in documentation.

1.4. Structure
- Section 2: Describes in detail any and all data used.
- Section 3: The methodology followed step by step.
- Section 4: Details on the analysis performed.
- Section 5: Reporting the results gathered.
- Section 6: Conclusions drawn from the project.
- Section 7: Further/Future research.
- Section 8: Appendices

2.0 Data

Due to the nature of the data used in this project several small datasets are used instead of larger datasets. Below is a brief description and overview of each dataset used before any pre-processing or transformation.

- "FormatShares" One of the larger datasets used within the project containing a breakdown of market share by music format i.e., CD, Vinyl, Streaming. This dataset was sourced from Kaggle.com and covers the years from 1973-2019.
- "SonyRevenue" Dataset containing revenue figures for Sony Music Group, covering the years 2006-2020. This dataset was sourced from Statista.com
- "UniversalRevenue" Dataset containing revenue figures for Universal Music Group, covering the years 2004-2020. This dataset was sourced from Statista.com
- "WarnerRevenue" Dataset containing revenue figures for Warner Music Group, covering the years 2004-2020. This dataset was sourced from Statista.com
- "SpotifyRevenue" Dataset containing revenue figures for the Spotify streaming platform covering the years 2013-2020. This dataset was sourced from Statista.com
- "AusIndustryRevenue" Dataset containing music industry revenue figures for the country of Australia, covering the years 2004-2019. This dataset was sourced from Statista.com
- "ChinaRevenue" Dataset containing music industry revenue figures for the country of China, covering the years 2014-2020 and predictions until 2023. This dataset was sourced from Statista.com
- "USRevenue" Dataset containing music industry revenue figures for the United States, covering the years 2009-2019. This dataset was sourced from Statista.com

- "EuropeRevenue" Dataset containing music industry revenue figures for all countries within the European Union, covering the years 1997-2019. This dataset was sourced from Statista.com
- "IndustryRevenue" Dataset containing worldwide revenue figures for all areas of the music industry, covering the years 1999-2019. This dataset was sourced from Statista.com
- "StreamingRevenue" Dataset containing revenue figures from worldwide streaming, covering the years 2005-2019. This dataset was sourced from Statista.com
- "AusPlatformSales" Dataset containing the number of digital and physical units sold in Australia from 2005-2018. This dataset was sourced from Statista.com
- "Big3Share2019" Dataset containing details on music industry company market share from 2019. This dataset was sourced from Statista.com
- "RevenueDistribution2019" Dataset containing details on revenue distribution in the music industry from 2019. This dataset was sourced from Statista.com
- "USChangeInSales" Dataset containing year-on-year percentage change in sales in the US market by format from 2009-2019. This dataset was sourced from Statista.com
- "ChinaForecasetUsers" Dataset containing details on the number of digital music users in China from 2017-2020 with predictions until 2025. This dataset was sourced from Statista.com

The below table gives more in-depth details on attributes, file type, file size, etc. for all datasets used

Data File	No. of Records	Description	Attributes	File Size	Data Format	File Type
FormatShares	8	Details of market share by music format from 1973-2019	1973,…,2019	3kb	Structured	csv
SonyRevenue	15	Sony Music Group revenue from 2006-2020	Year, Amount	1kb	Structured	csv
UniversalRevenue	17	Universal Music Group revenue from 2004-2020	Year, Amount	1kb	Structured	csv
WarnerRevenue	17	Warner Music Group revenue from 2004-2020	Year, Amount	1kb	Structured	csv
SpotifyRevenue	8	Spotify revenue from 2013-2020	Year, Amount	1kb	Structured	csv
AusIndustryRevenue	16	Australian music industry revenue from 2004-2019	Year, Amount	1kb	Structured	csv
ChinaRevenue	10	Chinese music industry revenue from 2014-2020 with predictions for 2021-2023	Year, Amount	1kb	Structured	csv
USRevenue	11	US music industry revenue from 2009-2019	Year, Amount	1kb	Structured	csv
EuropeRevenue	23	European music industry revenue from 1997-2019	Year, Amount	1kb	Structured	csv
IndustryRevenue	21	Total music industry revenue from 1999-2019	Year, Amount	1kb	Structured	csv
StreamingRevenue	15	Total streaming revenue from 2005-2019	Year, Amount	1kb	Structured	csv
AusPlatformSales	14	Details of the number of digital and physical units sold in Australia from 2005-2018	Year, Digital, Physical	10kb	Structured	csv
Big3Share2019	4	Details of music industry market share by company from 2019	Company, Share	9kb	Structured	csv
RevenueDistribution2019	6	Details of the revenue distribution in the music industry from 2019	Format, Share	9kb	Structured	csv
USChangeInSales	10	Details of change in US sales year-on-year by percentage from 2009-2019	Year vs. Year, Overall music sales, Total album sales, Digital album sales, Digital track sales, LP/vinyl album sales, CD	10kb	Semi Structured	csv
ChinaForecastUsers	9	Details of China's digital users from 2017-2020 with forecasts from 2021-2025	Year, Music Downloads, Music Streaming, Digital Music	10kb	Structured	csv

3.0   Methodology

Knowledge Discovery in Databases (KDD) was the methodology followed in this project. KDD was chosen because of its emphasis on research than a methodology like CRISP-DM. Knowledge discovery of changes in music industry revenue and consumption is the objective of this analysis and KDD fits the goals of this project and gives a good framework and approach to data analysis projects. KDD follows several key steps which are outlined and explained in the context of this project below.

3.1 Data Selection and Description

This step of the methodology focuses on the details surrounding the selection of data and a brief description of the data used. When setting out the purpose of the project some pieces of data were marked as being key to the project, this includes a breakdown of market share by format in the music industry, various pieces on revenue data like countries and companies as well as breakdowns in revenue by category like streaming and sales and some forecasts for music industry revenue figures to compare the forecasts carried out for this project. Each dataset selected can be classified as time series data i.e., a sequence taken at successive equally spaced points in time. In this case each dataset has values that progress year-on-year. Each dataset was downloaded in comma separated values (.csv) format which is the standard file type when downloading from Statista.com and apart from the USChangeInSales dataset, there no missing values in any of the data collected. Below is a table containing details on each file and how they are structured with file size and all attributes contained.

3.2 Data Processing and Transformation

After selecting all of the datasets for the project I then had to process, transform and combine all relevant datasets. As only the USChangeInSales data had missing values no data retrieval or removals had to be performed for all other datasets. After downloading the datasets from Statista each dataset contained 2 sheets when opened in MS Excel, the first sheet contained a link back to Statista and various other pieces of information that were not relevant to the project and interfered with the formatting when importing into R Studio so each one of these sheets within the datasets was removed in Excel so as they could be imported into R Studio correctly. Below is a list of processing and transformations carried out on all of the datasets with a table detailing the final output of compiled datasets.

1. When processing the "FormatShares" dataset the years 1973-2005 were removed from the dataset to ensure that the data left would only cover the period of time stated at the beginning of the project. When the dataset was imported into R Studio the attribute names all began with an X so for example it was "X2006, X2007, etc.", so each attribute name was changed, and the X was removed.

2. With the next several datasets being combined into one similar pre-processing and transformation was carried out on each of the following datasets
   a. "EuropeRevenue" – Removed values from the years 1997-2013 and renamed the year column from "ï..Year" to "Year"
   b. "AusIndustryRevenue" – Removed values from the years 2004-2013 renamed the year column from "ï..Year" to "Year"

c. "USRevenue" – Removed values from the years 2009-2013 renamed the year column from "ï..Year" to "Year"

d. "ChinaRevenue" – Removed values from the years 2020-2023 as we do not need the prediction values renamed the year column from "ï..Year" to "Year"

It is important to note each dataset was reduced to only the years 2014-2019 as this was the only common data between the 4 datasets so this was used for the analysis. Each dataset's "Amount" attribute contains the amount by million US$ so that each region is measured on a common range. The data was combined into a new data frame and written to the file "RegionRevenues.csv"

3. The next 4 datasets were also combined into one combined dataset so again similar pre-processing and transformation was carried out on each dataset
   a. "SonyRevenue" - Renamed the year column from "ï..Year" to "Year"
   b. "WarnerRevenue" - Renamed the year column from "ï..Year" to "Year" and removed values for 2004 and 2005.
   c. "UniversalRevenue" - Renamed the year column from "ï..Year" to "Year" and removed values for 2004 and 2005.
   d. "SpotifyRevenue" - Renamed the year column from "ï..Year" to "Year"

While the first 3 datasets in this collection cover the years 2006-2020, the Spotify data only covers from 2013-2020 as Spotify was launched in 2008, they only began to gain traction in the early 2010's. Each dataset in this collection's "Amount" attribute is also the amount by billion US$. The data collection was then combined into 2 datasets, the first contained only the first 3 datasets which was written to the file "CombinedRevenue.csv", with the second containing all of the datasets written to the file "CombinedRevenueWithSpotify.csv".

4. The next pair of datasets were being combined into one dataset and were pre-processed and transformed in similar fashion.
   a. "IndustryRevenue" - Renamed the year column from "ï..Year" to "Year" and removed values for the years 1999-2004.
   b. "StreamingRevenue" - Renamed the year column from "ï..Year" to "Year".

Both of these dataset's amounts were again the amount by billion US$ and were written to the "IndustryVsStreaming.csv" file.

Details of all of the combined datasets are contained in the table below.

Data File	No. of Attributes	Attributes	Data Type
RegionRevenues	5	Year, EuropeRevenue, AusRevenue, USRevenue, ChinaRevenue	Numerical
CombinedRevenueWithSpotify	5	Year, SonyRevenue, UniversalRevenue, WarnerRevenue, SpotifyRevenue	Numerical
IndustryVsStreaming	3	Year, IndustryRevenue, StreamingRevenue	Numerical

All other datasets collected were ready for use after the sheet previously mentioned was removed from the file. These datasets were mainly used for supporting data and statistics.

3.3 Data Mining

Because of the nature of the time series data and the small size of the datasets I consulted with my lecturers about what kind of data mining could be carried out on the collected data and I was recommended using a Holt-Winters method on the data. The H-W algorithm is used for forecasting time-series data and generally gives a more accurate prediction than moving averages as it follows the linear trend of the data to predict over the given timeframe. This test was performed in SPSS using the time series modeler function and visualisations were generated so the trend could be interpreted.

3.4 Interpretation/Results

When extracting results from the data visualisations were the main method used for extraction. SPSS and Excel were all used for generating initial sample visualisations from the data, with PowerBI being used after this to generate more refined and attractive visualisations for results analysis as well as using them in documentation.

3.5 Implementation

From the beginning of the implementation of the methodology, any files that were required to be processed or combined were imported into R Studio to do so. The R programming language as well as packages within R Studio were used, with the main package used being the tidyverse package. This package is a compilation of other packages like dplyr, tidyr and ggplot2 as well as many others that make cleaning, processing and visualising of the data extremely easy. Functions like rename() and cbind() were some of the most frequently used when processing and transforming the data. Excel was also used for some data cleaning were there were multiple sheets in the csv files that were irrelevant to the analysis as well as viewing the raw data after datasets were combined and written to new csv files.

SPSS was used when performing the forecasting for the relevant data using the time series modeler. The time series modeler in SPSS allows to you to pick the exponential smoothing method and then the Holt's linear trend criteria which is what is needed to generate the forecast. After this the forecasts were saved and exported from SPSS to a new csv file so that a graph can be plotted to analyse the new data.

PowerBI, SPSS and Excel were all used to generate visualisations but PowerBI was used for the visualisations used in this report as well as the ones used for results analysis as it an industry standard analysis tool and I have previous experience with the software from my 3rd year internship so I knew how to generate easy to read, high quality visualisations from the software that were all analysed to draw results from that were then reported in this document. The graphs and plots used in this project were pie charts, histograms, stacked bar charts as well as stacked area charts.

4.0   Analysis

The objective of this project is to analyse how various revenue streams in the music industry have changed in the last 15 years so each of the datasets detailed in section 2 will be used in
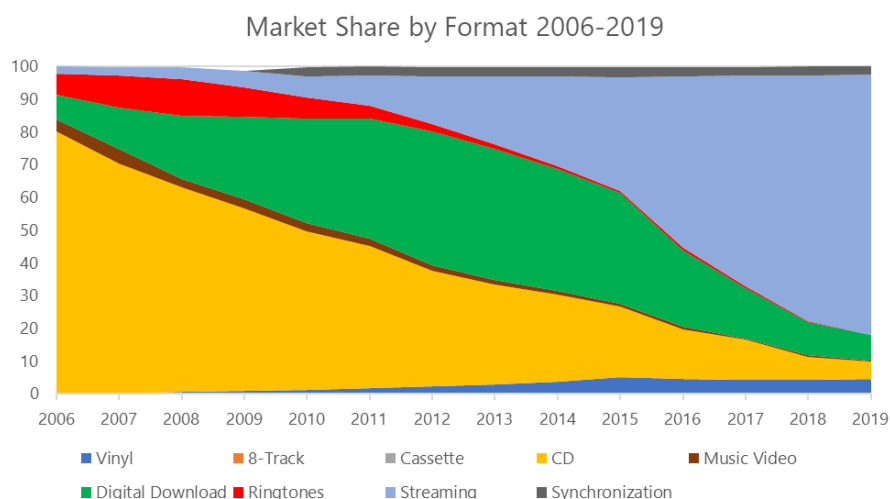
the analysis for this research with some of the data being used in several of the results and some being used as supporting data for the main results. With the data being majority revenue and sales data, visualisations were the most used method of analysis as it is easy to assess any trends in the data over the given time period as well as the format market share dataset. With the datasets being so small and also having many of them there weren't many other methods available to me when analysing the results but this lets the end user see the results exactly as I did when compiling them. When conducting the forecasting tests for each of the relevant datasets the dependent variable in each case was the data column like "SonyRevenue" or "IndustryRevenue". After the tests returned results and they were exported as required, more visualisations of this additional data were generated and analysed against the original data to see how the data is predicted to trend over the three year time period set out.

5.0   Results

Below I will discuss each of the results that were drawn from the data, there are four results in total with aspects of data used in multiple areas.
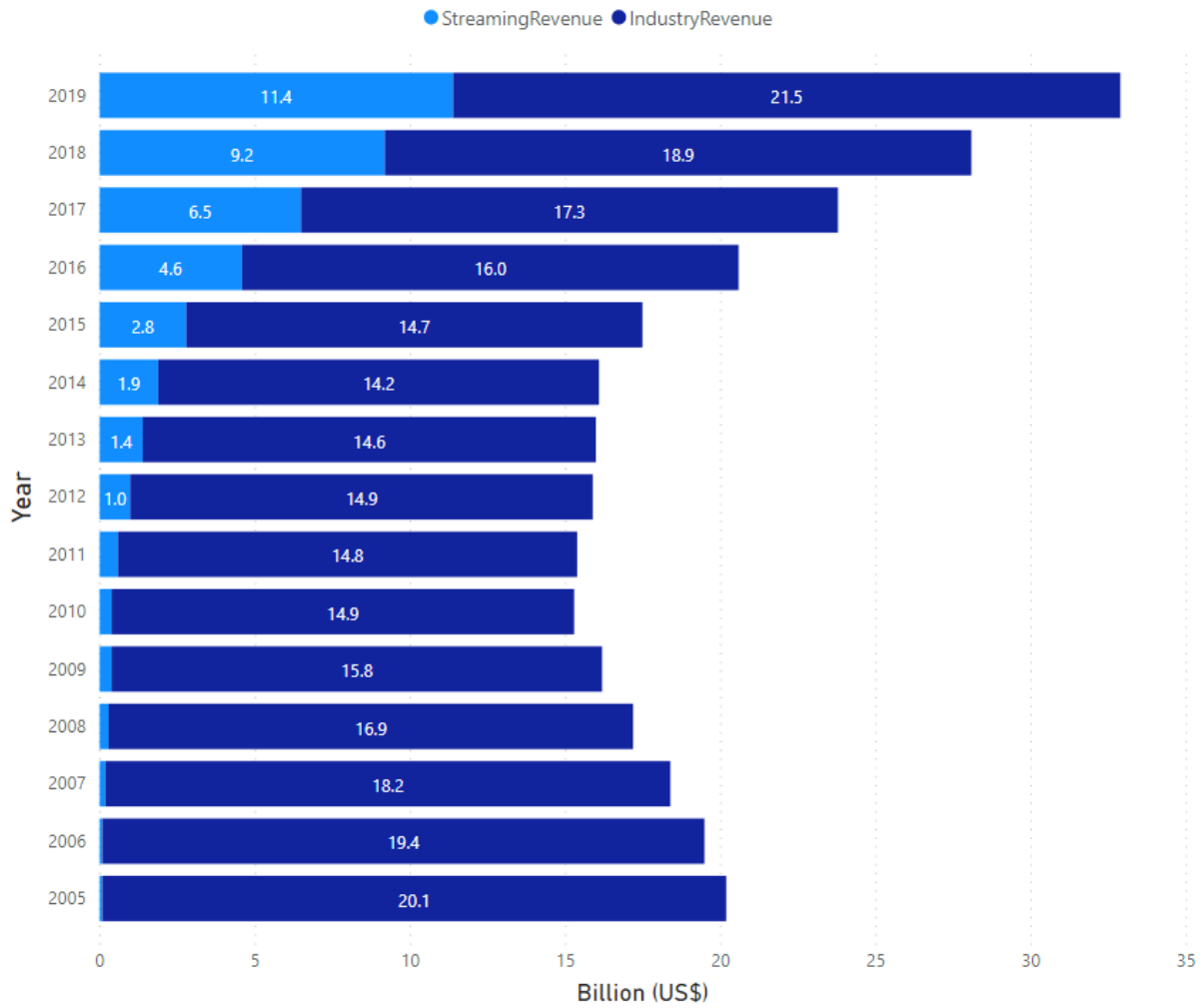
5.1 How Has Industry Revenue Changed Since the Popular Adoption of Streaming

In the below graph we can see the last 14 years and how market share has progressed from 2006-2019. From 2006 the market share for CD's was at it's peak of almost 80% but steadily dropped as the years went on, only having a tiny portion of market share by 2019. On the other side of the coin we can see that while streaming has been on the rise since 2006 it begins to rise rapidly from 2013/2014 onwards and reaches almost 80% of market share by 2019. Another significant player in this graph is Digital Downloads which enjoyed a sizeable portion of market share from 2009-2016 but ultimately fell off due to streaming being a cheaper alternative to digital music consumption.
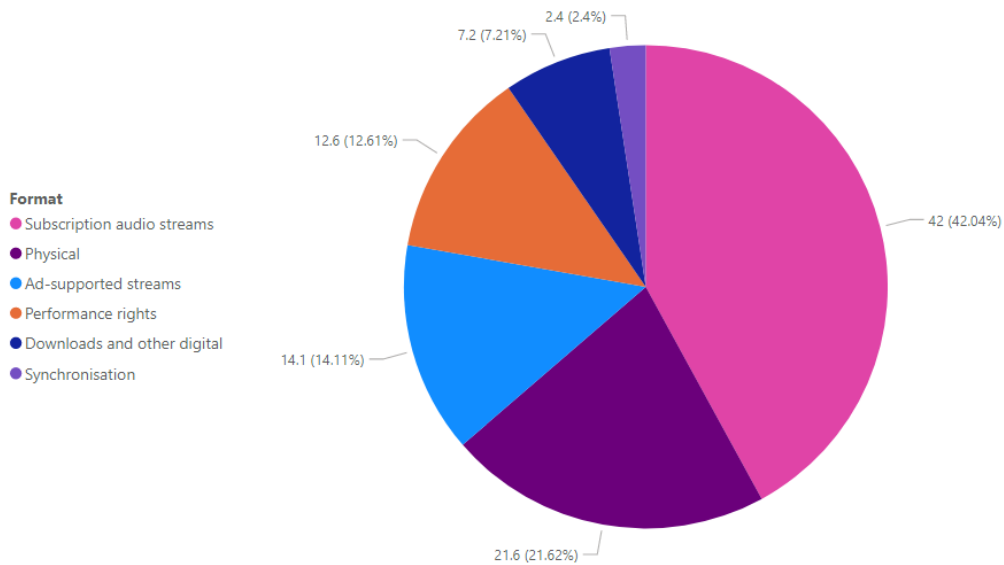

Market Share by Format 2006-2019

When assessing total music industry revenue in the below plot we can see industry revenue begins in 2005 at $20 billion and steadily declines before plateauing and then ultimately rising and surpassing it's high from 2005. The fall in revenue cannot solely be attributed to CD's waning sales however as around that time period there were a number of outside factors like economic recession that likely had an effect on overall sales.

Total Music Industry Revenue and Total Streaming Revenue by Year

● StreamingRevenue ● IndustryRevenue

Year	StreamingRevenue	IndustryRevenue
2019	11.4	21.5
2018	9.2	18.9
2017	6.5	17.3
2016	4.6	16.0
2015	2.8	14.7
2014	1.9	14.2
2013	1.4	14.6
2012	1.0	14.9
2011		14.8
2010		14.9
2009		15.8
2008		16.9
2007		18.2
2006		19.4
2005		20.1

Billion (US$)

Above we can see the revenue for streaming which has made massive gains year-on-year since 2005 and now accounts for more than half of all music industry revenue. This is interesting as while it dominates market share, streaming is also not a very big earner as a typical monthly payment would be €10 which would be cost of just 1 CD or a portion of the price of a vinyl record. This is further broken down in the below pie chart that breaks down the distribution of worldwide music industry revenue for 2019.

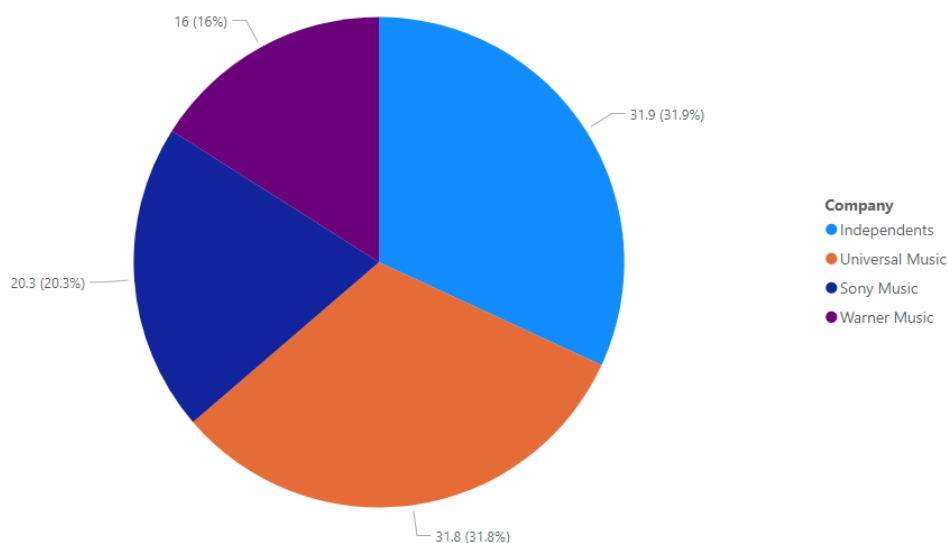Music Industry Revenue Distribution for 2019

As can be seen while the revenue from subscription audio streams accounts for nearly 40% of all music industry revenue it had almost 80% of market share as seen in the first graph. We can conclude from this information that while almost 80% of music is consumed via streaming, this does not necessarily translate over to equal revenue distribution in the industry.
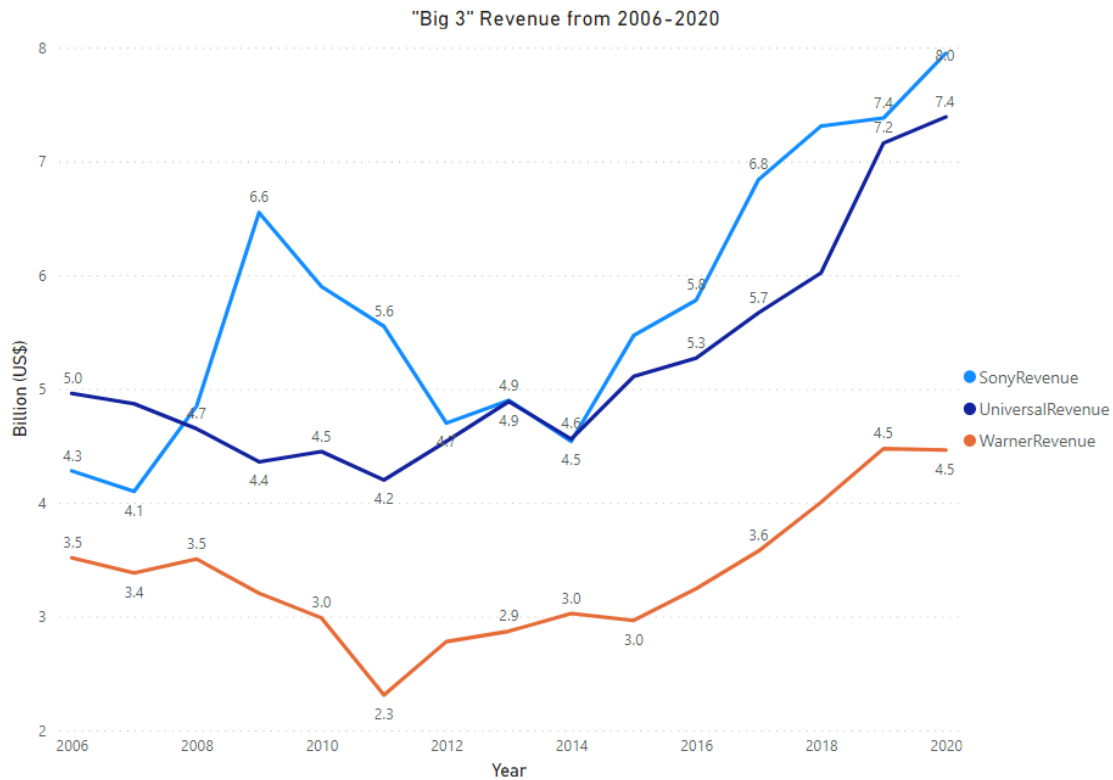
5.2 How Does Spotify's Revenue Compare to the "Big 3"

When looking at the music industry there are the "big 3" companies that the majority of major artists are signed to, with some artists being directly signed to these labels or a subsidiary of these larger labels. These three companies control the majority of market share in the music industry and as seen below, in 2019 they control almost two-thirds of the music market.
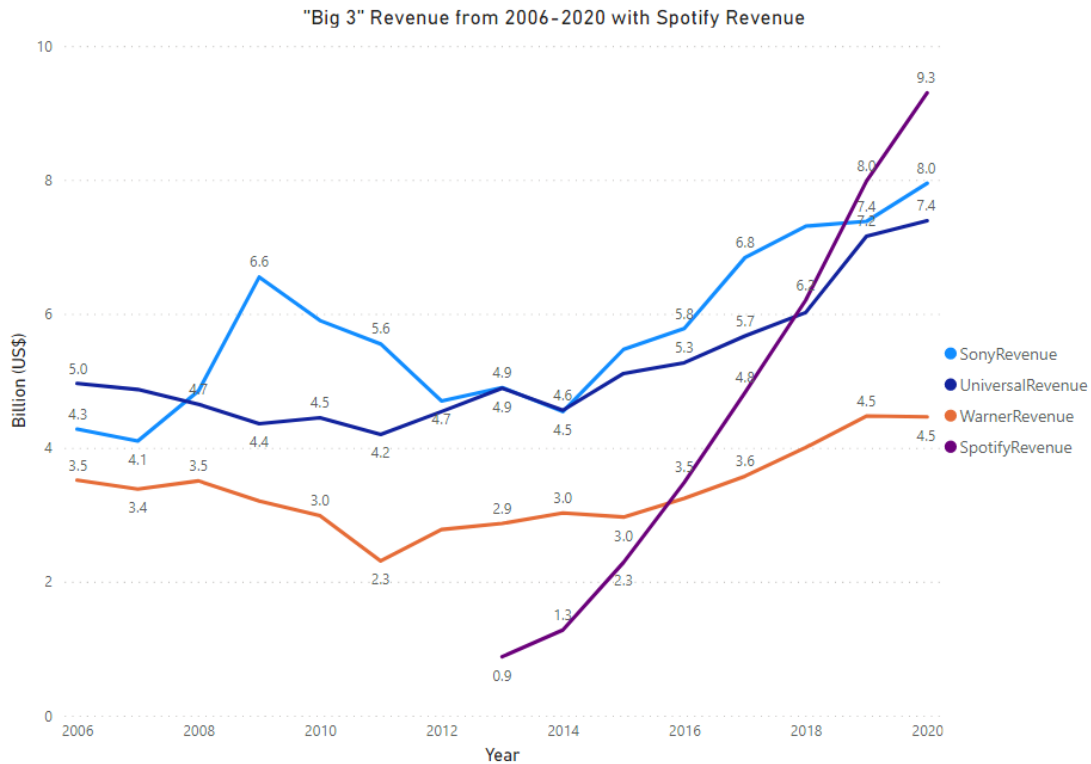
Global Market Share by Company

These companies make their money mainly from record sales/deals, performance rights and various other avenues from these artists and is made up of Warner Music Group, Sony Music Group and Universal Music Group. All three companies are based in the US and control a large part of the music market due to their size and wealth. Below is a plot with all 3 companies revenues compared over the period 2006-2020.
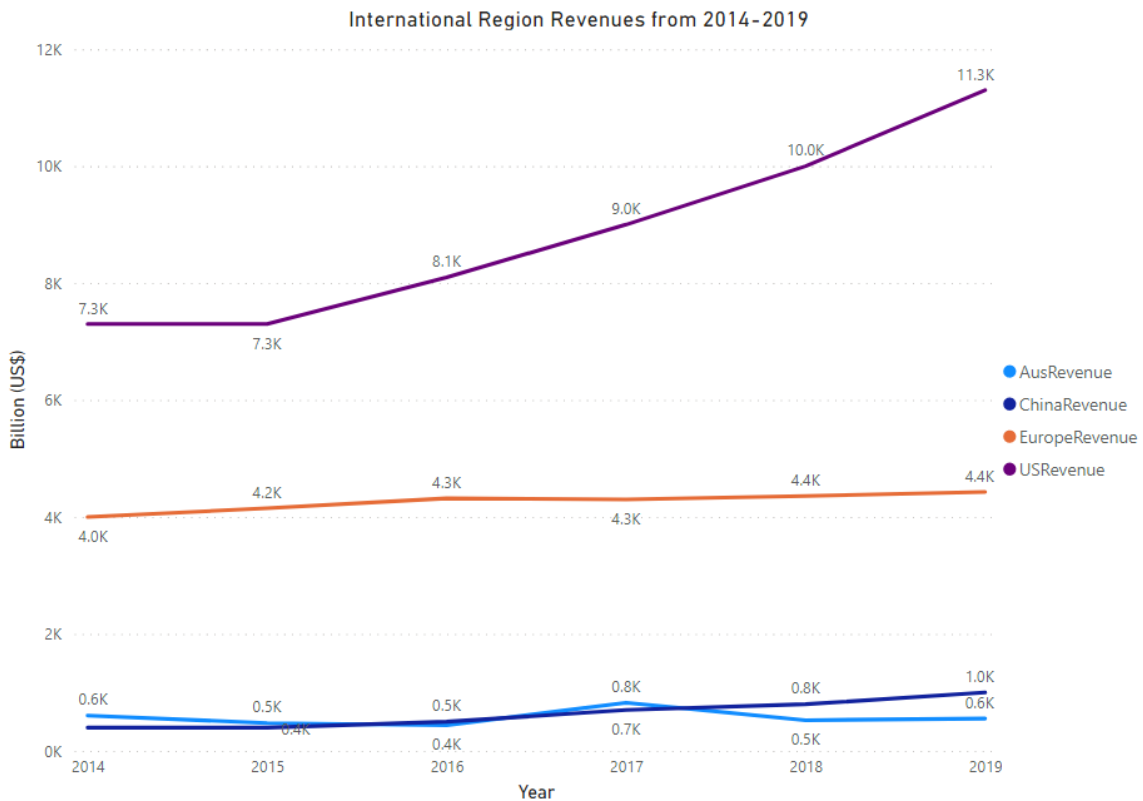

As we can see each company follows a different trend over this time period with Sony and Universal enjoying more success than Warner in recent years. Interestingly all 3 company's revenue begins to climb from 2014 onwards and reaches their peaks in 2020 with Sony coming out on top with $8 Billion in revenue and Universal being close behind with $7.4 Billion and Warner being behind with a not-so-little $4.5 Billion. Universal and Warner follow a relatively similar trend but with Warner's lows being a bit more dramatic than Universal's, with the latter having more dramatic highs with more overall revenue. As mentioned in 5.1, there are outside factors like economic recession that likely resulted in revenue stagnation for Warner and Universal from 2006-2014 with Sony initially performing well but ultimately left with a significant drop in revenue for several years until they surpassed their 2009 high of $6.6 Billion 10 years later with $6.8 Billion.

"Big 3" Revenue from 2006-2020 with Spotify Revenue

So where does Spotify fit in? As the world's largest music streaming service their revenue is comparable to the "big 3" in recent years even though they have business dealings with all 3 companies. As we can see in the above plot over the common period of 2013-2020 Spotify's revenue steeply rises from being significantly lower than the big 3 to surpassing them inside 7 years. This shows that as streaming and digital services became more widely adopted worldwide, the revenue of the biggest service has surpassed that of the biggest record labels in the industry in such a short space of time and may continue to rise exponentially as it's user base expands and digital domination in the market continues. This data does not take into account other streaming services like Apple Music or Soundcloud which are also "freemium" music streaming subscription platforms and are Spotify's largest competitors in the space.
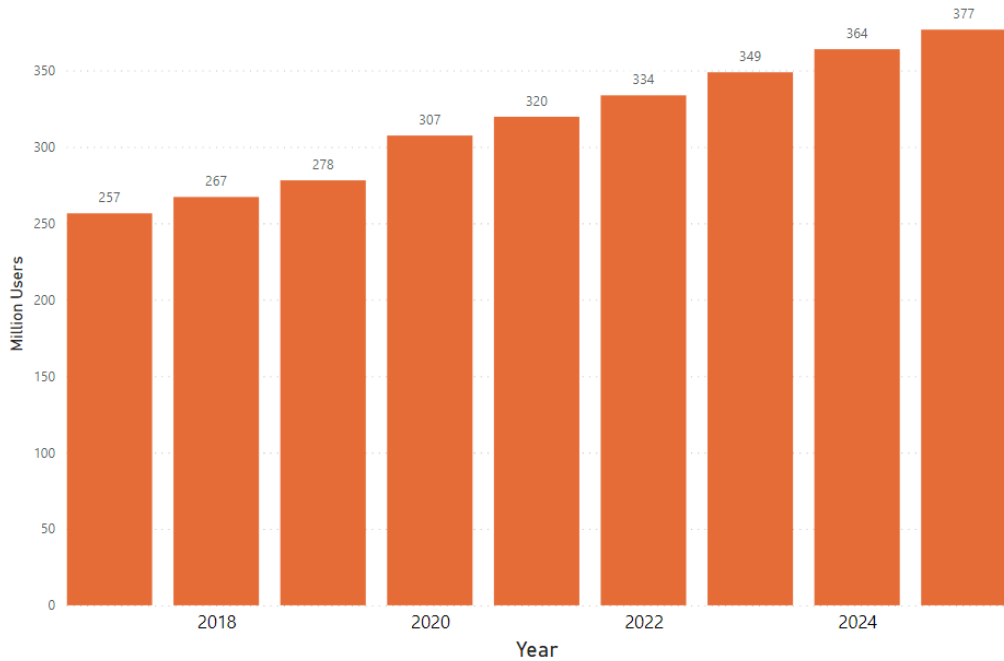
5.3 Comparing International Industry Revenue

At the moment the Australian, Chinese, European and US markets make up some of the biggest international markets in the music industry but all vary in the amount of revenue they pull in when looking at recorded music industry revenue. Recorded music industry revenue is any revenue made up of record deals and music sales, whether digital or physical. Below is a plot comparing each of the aforementioned markets revenue over the 2014-2019 period.
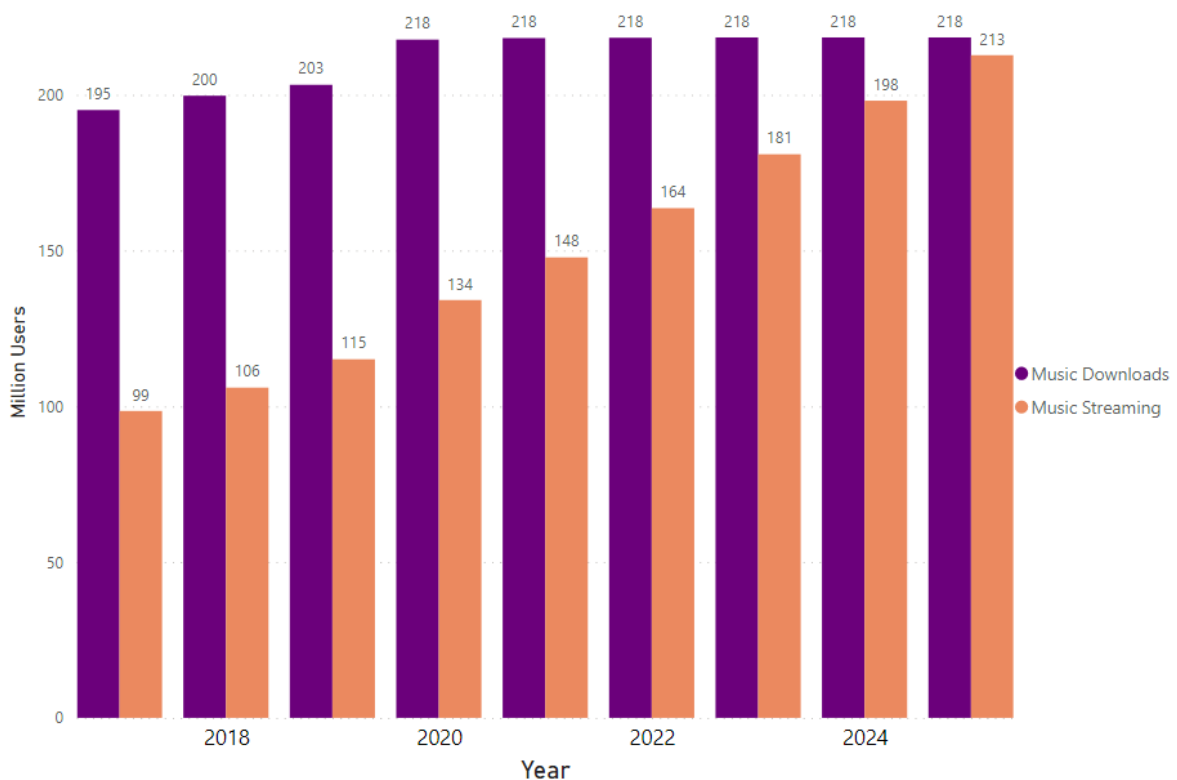
International Region Revenues from 2014-2019

As could be guessed the US has the largest market of the four followed by Europe with China and Australia wrestling for third place. Even though the US has a far smaller population than China or Europe it's recorded music industry revenue is highest due to having the "big 3" music companies based there as well as a number of large international music labels/distributors based there. Although Europe has quite a large population when it comes to the recorded music industry it is quite self-contained. A large factor in this is Europe's make-up, i.e. it is made up of a number of countries with different cultures, languages and music styles that do not necessarily translate country to country whereas the US is one singular market, the EU is made up of various smaller markets. This results in various smaller music companies/labels/distributors being based in Europe and hence the lower total revenue than the US. Down at the bottom of the plot we can see Australia and China battling it out over higher revenue. It may seem strange that a country like China with such a large population is pitted against Australia when it comes to recorded music industry revenue but due to censorship of a lot of western media and a small home market China is lagging behind the leaders, though many do believe China could become one of the fastest growing markets in the future as the industry finds it's feet there. With the below predictions by Statista estimating that China's digital music users could grow to 376 million by 2025.

Chinese Digital Music Users 2017-2020 with Forecasts (2021-2025)

However, this prediction is based on digital users and hence digital medium's like streaming and digital download will likely not have a huge impact on recorded music revenue as the below graph shows that users that download music has stagnated for several years while streaming users has steadily risen in the same time.

Chinese Digital Music Users 2017-2020 By Format with Forecasts (2021-2025)

On the other side of the world in the US, the table below shows the performance of various formats across a number of years. While the digital sales initially increase year-on-year, they begin to join CD's in falling in sales, coinciding with the rise in streaming around this time. Interestingly across all of this time the only format to rise in sales is Vinyl records. This is the only physical medium to continually increase in sales is due to their high quality sound and the recent resurgence in "vintage" commodities.

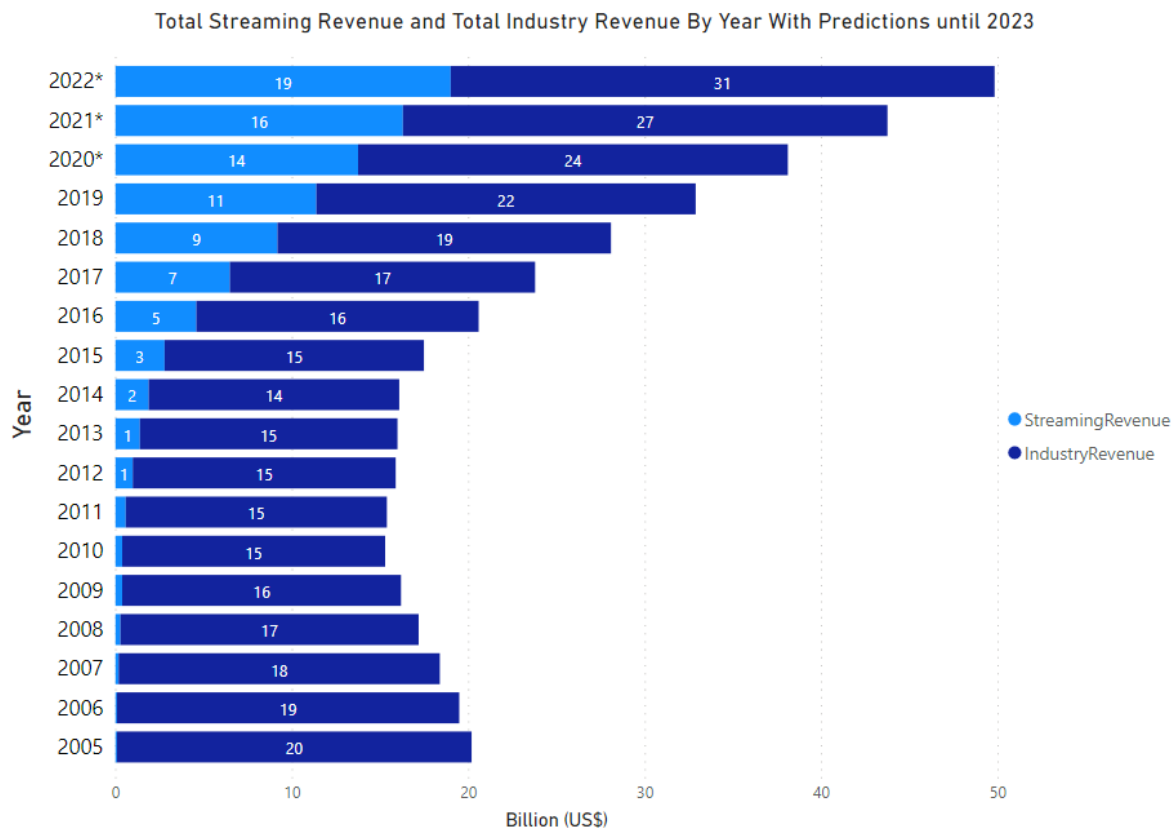| Year vs. Year	Overall music sales	Total album sales	Digital album sales	Digital track sales	LP/vinyl album sales	CD	
2010 vs. 2009	-2.4	-12.7	13	1	14	-	in %
2011 vs. 2010	6.9	1.3	19.5	8.5	36.3	-	in %
2012 vs. 2011	3.1	-4.4	14.1	5.1	17.7	-13.5	in %
2013 vs. 2012	-6.3	-8	0	-6	33	-14	in %
2014 vs. 2013	-	-11.2	-9.4	-12.5	51.8	-14.9	in %
2015 vs. 2014	-	-6.1	-2.9	-12.5	29.8	-10.8	in %
2016 vs. 2015	3.1	-16.7	-20.1	-25	10	-16.3	in %
2017 vs. 2016	12.5	-17.7	-19.6	-23.4	9	-	in %
2018 vs. 2017	-	-17.7	-20.7	-27.2	14.6	-20.9	in %
2019 vs. 2018	-	-18.7	-23.5	-25	14.5	-	in %

Down Under we can see Australia's digital and physical sales in the clustered column chart below. As is the trend when looking at the results from 5.1 physical sales start off strong and continually falls off as time progresses while digital sales are inverted, starting low and rising to dwarf physical sales by 2018. Surprisingly while digital sales are significantly higher than physical sales by 2018, they are still not as high as the peak of physical sales reached in 2005 of over 500 million units.

Australian Sales by Format 2005-2018

5.4 Forecasting for the Future
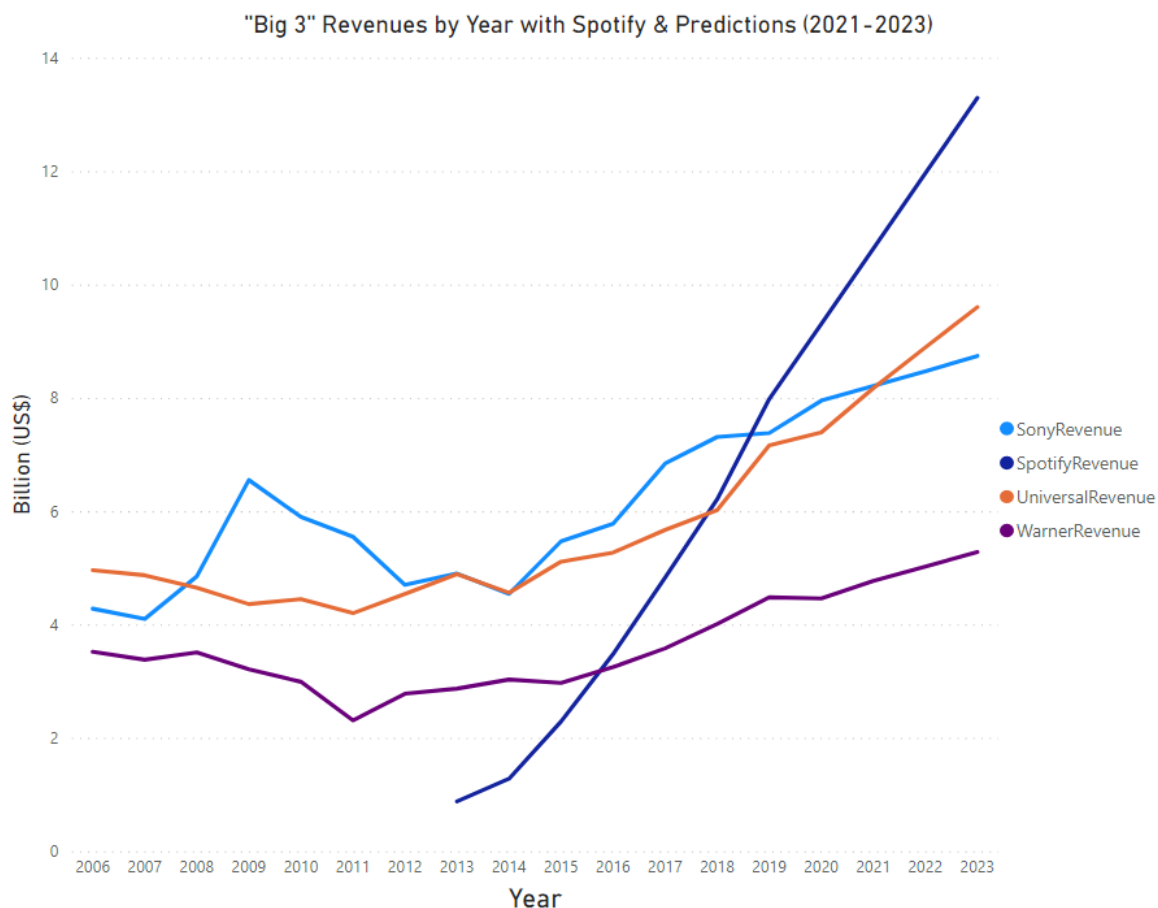
5.4.1. Industry Vs. Streaming

So based on the data from 5.1 we are forecasting total music industry revenue versus total streaming revenue for the following three years, the chart for which can be seen below.

Total Streaming Revenue and Total Industry Revenue By Year With Predictions until 2023

So as we can see both industry revenue and streaming revenue is forecasted to continue steadily increasing until 2022 using our method. By 2022 streaming revenue has been forecasted at $19 Billion and total industry revenue at $30 Billion so by 2022 Streaming revenue will make up almost 66% of total music industry revenue according to our forecasting model. As there is unlikely to be a change back to CD's and other physical formats soon and with the diminishing revenue of digital downloads this forecast of streaming making up a majority of total music industry could likely come to fruition at some point whether it is 2022 or several years beyond.

5.4.2 "Big 3" Vs. Spotify

From the results of 5.2 when comparing Spotify's revenue vs. the "big 3" music companies we have continued on and made forecasts from the data as seen below.

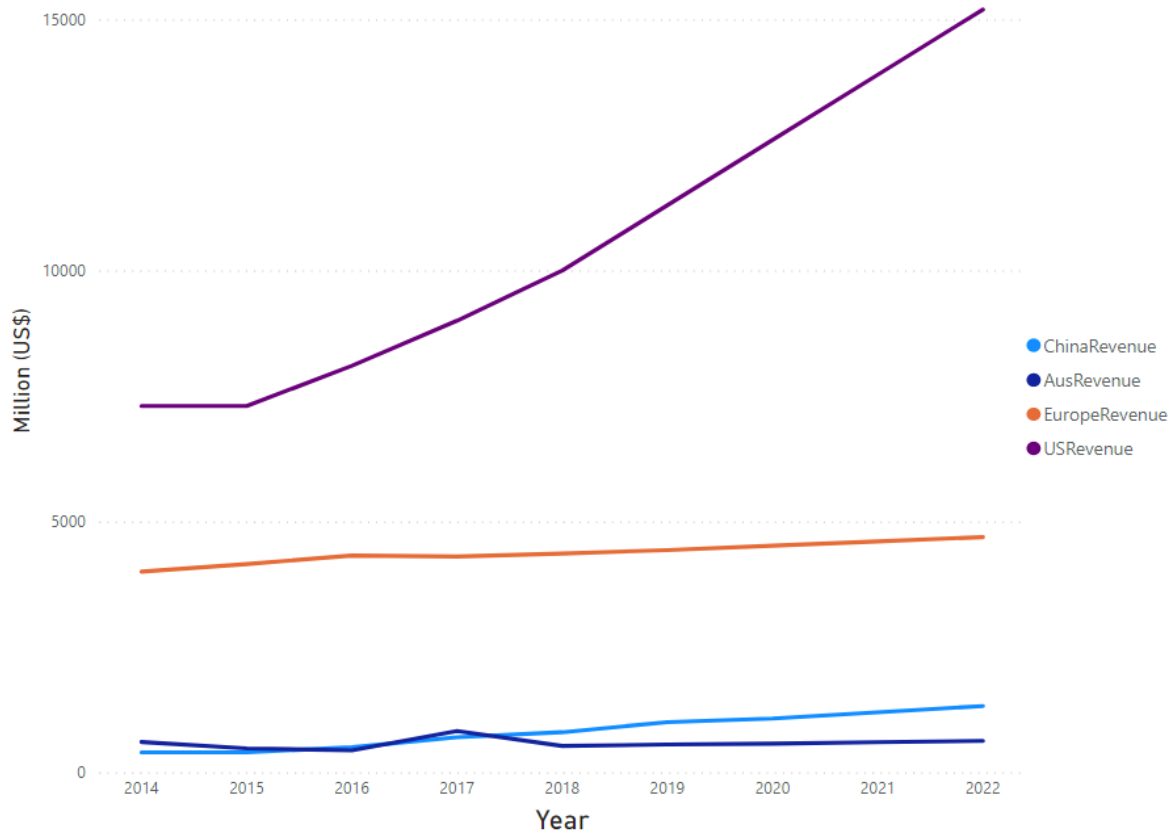"Big 3" Revenues by Year with Spotify & Predictions (2021-2023)

So from the chart we can see all four companies revenue's are expected to continue their upwards trend over the next three years. The "big 3" companies continue to increase gradually over the forecasted period and look to make solid gains but Spotify outshines them in the chart, continuing their sharp increase to finish at just over $13 Billion by 2023. This puts them almost $4 Billion higher than the closest company (Universal) and well above Sony and Warner. This forecast obviously cannot take into account how the pandemic has affected revenue for these companies as a case could be made that their revenue may be increased as more people may be listening to music at home or that people may not be consuming as much music as before due to less radio listenership which is a huge income for the big 3. It will be interesting to see how accurate these forecasts turn out to be over the next three years and assess how the pandemic has affected each of the company's revenue.

5.4.3 International Recorded Music Industry Revenue
Based on the results and data from 5.3 we have made forecasts on the four regions from 2020 to 2022. Unfortunately there was no data available for 2020 online so forecasts had to be made for that year as well.

International Regions Music Revenue by Year with Predictions (2020-2022)

As can be seen above the forecasts all continue in the same way as the previous years with Europe and Australia rising slightly, China rising a little higher and the US continuing to rise sharply over the coming years. With the US projected to have just over $15 Billion revenue in 2022, this would correspond to 50% of total music industry revenue as forecasted in 5.4.1. This is a significant amount of revenue control for the US and if the forecasts do turn out to be somewhat accurate this could be the start of the US completely dominating the international music industry financially. While the US is making huge gains, the remaining regions are predicted to plateau and make minimal increases to their revenue.

6.0   Conclusions

From the gathered results we can see how the industry as a whole has moved towards digital music consumption with streaming being at the forefront of this digital wave. We can conclude that the move digitally has coincided with an overall increase in revenue, but as the old saying goes, correlation is not causation and that phrase may apply in this case. While digital music consumption is the most common form of consumption today, physical methods still hold a good portion of consumption with older audiences as they may struggle to keep up with technological advances and stick to their current methods of listening.

One of the advantages of this research project is that it gives an insight into various aspects of the music industry; who are the main players, how money is made and how music is listened to, and it is covered around the world and over a common timeframe that is easy to

follow and could be of interest to a wide range of people that could be interested in music or consumer behaviour or corporate performance. There are some drawbacks to this project like the lack of insight on a smaller level like country by country or how smaller independent music labels perform versus their larger competition. This granularity is hard to capture as it is quite difficult to find it compiled into datasets that are available online but I think the general approach to various areas that was taken captures enough detail to give an insight into each respective area and the performance for each as well as the predictions for the future.

7.0   Further Development or Research

If I were to progress this project in my own time with no time constraints I would try to reach out to various government bodies that have collected more in-depth data on the music industry in their own respective countries and compile and compare the data on a country by country basis with more detail in the data and larger datasets. Part of the limitations of this project was the smaller datasets that really held me back on using more analysis techniques. Another aspect I would examine if I were to redo this project would be covering a wider timeline, this would make data collection slightly more complicated as the data going back to perhaps the 1980's and before would not be as in-depth as any data collected in modern times but it would still be interesting nonetheless and the "FormatShares" dataset before transformation could be used as this covers from 1973-2019. Looking into the future I think it would also be interesting to come back to this project in three years to examine how accurate the predictions made were and how they could be adjusted based on their performance to get better predictions for the following three years as well as examine how the pandemic impacted the music industry and how this may have skewed any trends in revenue. There are various other branches that could be explored like Spotify listenership data and billboard charts data that I would be interested to do some research in and perhaps cross reference it with some work that has been done in this research and also see what other studies and research that is out there that could lead to some inspiration into a different avenue as this was an enjoyable project to undertake because of the personal interest.

8.0   Appendices

8.1. Project Plan

Parent Task	Sub Tasks	October	November	December	January	February	March	April	May
Research	Initial Research	■	■						
	Dataset Reseach		■	■					
	Midpoint Presentation			Dec 22nd					
Data Gathering	Data Gathering			■	■	■	■		
Data Cleaning	Cleaning and Sorting of Raw Data						■		
	Representing Data						■		
Analysis	Conducting Analysis of all data that's been gathered and cleaned						■	■	
Reporting	Reporting findings from the analysis							■	■
	Publishing								■
	Final Documentation, Presentation and Showcase								■

8.2. Reflective Journals

9.2.1 October

Now that we're one month in I've adapted to the online learning environment and am now familiar with all of my modules and lecturers. At the start of the semester I was having some difficulty in deciding what to propose for my project idea as I had no idea what a data analytics project actually entailed. I went back and forth on different ideas and eventually decided that it was best to work on a project that I had actual interest in, and not just a topic for the sake of it, so I eventually decided on my idea of analysing the breakdown of revenue in the music industry in the last 10-15 years and how COVID-19 has affected the industry this year. I settled on this because I have a good interest in music and I thought it would be interesting to see how the change in medium for music i.e. CD, Vinyl, Streaming, has changed in the last decade and how it's been affected this year. The modules that will definitely help most will be Business Data Analysis and Data Application Development as these are the modules most relevant to my project and I will be paying close attention over this semester so that I can apply any material learned to better analyse my project and report my findings.

9.3.2 November

This month I have spent time doing research into other projects on the music industry and other data analysis projects. I found the past projects on the NCI library very helpful just to get an idea of what the level and format required as well as what kind of timeline those projects were on. We had to do a data analysis for the data application development module and I definitely learned some new skills and techniques that will be very helpful for my project like using the R language in practice as well as packages within R Studio like ggplot that displays graphs and plots to display data. With the midpoint presentation being due for the 22nd of December I have started looking at getting a head start on completing some sections of that before I resume work on some of my other CA's/Projects that are due around the same time.

9.3.3 December

This month I mainly focused on my midpoint presentation which was due on the 22nd or December, I presented all of the work and research done so far as well as highlighting my plan over the coming months to keep on track and get everything done on time. December was a very hectic month with a lot of projects and assignments due around the same time as well as having terminal based assessments to be completed but I did learn about some useful technologies like MapReduce and how to use cloud services within the Data Application Development module. These technologies could turn out to be useful later in the project when analysing the data to find some new information within any data that has been cleaned and processed and python and AWS could potentially be some other technologies used within the project to add to the other technologies I have already highlighted in my midpoint presentation.

9.3.4 January

This month we started our final 2 modules of the year, Data and Web Mining and Advanced Business Data Analysis. Both modules are going to cover a lot of information on the steps in our projects that we haven't covered yet like doing some data mining on our final data to uncover some patterns as well as some more advanced analysis techniques not covered in semester 1. As according to my project plan, I am now in the data collection stage of my project. This stage should take me most of February and March to complete but due to the steps involved in this stage having already been covered in semester and some datasets being bookmarked I have a good idea of what to do and where to look for suitable data.

9.3.5 February

This month I started to combine some of the data I have collected into a number of smaller datasets. Due to the type of data I am collecting i.e. different income data from sections of the music industry I cannot combine all of my data into one large dataset instead I have to have several smaller datasets made up of similar data so that I can perform analysis on them and generate tables, graphs and charts from all of the individual datasets. I have been using R Studio to combine/transform the data in csv files which can be accessed in Excel and are also stored in MySQL. For the month of March I plan on finishing up my data collection and moving on to transformation and analysis stages.

9.3.6 March

This month I started to conduct my analysis and get results from my data. I calculated some moving averages in Excel and used SPSS for generating graphs and charts. I have been asking my data and web mining lecturer about what kind of techniques I can apply to my data as my datasets are all small in size but I am still waiting on a response on what to apply. At the moment I have used moving averages to predict some revenue figures for the future. I still need to find some tests to run on my data to add some complexity to my project which I will discuss with my lecturers/supervisor.

9.3.7 April

For the final month of working on my project I have gotten a head start on writing the report and cleaning up any files/code well before the final submission so I can put more focus on my presentation, I am planning on using R markdown for presenting some of my R code just for

a display of some of the work involved in the project. I will work to finish off the report after all of the TABA's and other assignments are out of the way as I will have some more free time to get everything in order before the submission date.

8.3 Project Proposal

National College of Ireland

Project Proposal

An Analysis of the Shift in Music Industry Revenue Streams in the Last 15 Years and Where 2020 Fits In.

6/11/2020

BSHC4

Data Analytics

2020/2021

Carl Flynn

X17347726

X17347726@student.ncirl.ie

Contents

Objectives

The objective of this project is to analyse the shift in revenue streams in the music industry in the last 15 years for both artists and labels alike. With streaming being the main medium in which music is consumed for the last few years I want to analyse the transition from physical sales (CD, Vinyl, etc.) being the main source of income for the industry to another source whether it be streaming royalties, live shows or otherwise. Following on from this I want to analyse how 2020 has impacted artist income with Covid-19 whether negatively or positively in some areas.

Background

I came up with this idea when I was thinking of how I could apply my own interests to this project as I don't want to spend so much time on a topic I have no interest in and I thought music would be a good avenue to explore as there is a lot to be analysed and upon doing some research I thought revenue would be a good area to analyse and I think analysing how Covid-19 has affected the industry is a very relevant topic today. While researching I found a number of good articles from Statista, The Huffington Post and The Guardian on how revenue has changed in the last 10 years but I thought a bigger timeframe would give a broader picture of just how much it has changed in a relatively short time. There have been a number of studies on how big data is being used in the music industry, how Covid-19 has affected the arts sector in the UK/US and a lot of statistics on revenue numbers in the last few years country by country as well as worldwide, all of these are relevant to my project and have given me insights into each respective area. As far as data goes I found some very useful datasets on Kaggle on music industry sales for different mediums i.e. cassette, vinyl, CD, streaming. For the last 40 years, while this data is useful for overall numbers I will need more in depth numbers with further breakdowns which are also available online. I will mostly be using secondary data as it would be difficult to compile all of this data myself. I think this project will be interesting to the average person as it can show just how much they're favourite artists may have been affected this year as well as give an insight into just how much money is in the music industry as a whole.

Technical Approach

At the moment and for the coming weeks I am/will heavily research other data analytics projects, any relevant journals, articles, etc. and see what the kind of standard I am aiming for is, after this I intend to research more into data gathering methods as well as apply what I've learned in the Data Application Development module to begin my data gathering. We have been learning R so far and are moving onto Python soon and I'm sure they will be very useful tools for my project. As I intend to use mostly secondary data I will gather datasets online from various sources as well using some web scraping and web mining as well as API's to extract data. Once all my data is collected I will then clean up the data, removing any duplicate entries, irrelevant information, etc. JASP, Rattle or Orange may be useful tools to assist with this. Once all these steps have been complete I will move on to my exploratory analysis with the data and draw my conclusions and report my findings in my final report.

Special Resources Required

N/A

Project Plan

Parent Task	Sub Tasks	October	November	December	January	February	March	April	May
Research	Initial Research	■	■						
	Dataset Reseach		■	■					
	Midpoint Presentation			Dec 22nd					
Data Gathering	Data Gathering			■	■	■	■		
Data Cleaning	Cleaning and Sorting of Raw Data						■		
	Representing Data						■		
Analysis	Conducting Analysis of all data that's been gathered and cleaned						■	■	
Reporting	Reporting findings from the analysis							■	■
	Publishing								■
	Final Documentation, Presentation and Showcase								■

At the moment I am still researching other data analytics projects and dataset researching. We were informed that the gathering of data will be taking the majority of the time in the project lifecycle so I have given the most time to this. The rest of the timeline I hope to stick to and I feel I have given generous time to complete.

Technical Details

R, Python – Will be used for analysis of the data in R Studio.

Kaggle, Google Datasets, Statista – Will be used to retrieve datasets.

API's – Will be used to retrieve data from various sources, example: Spotify API.

Evaluation

Load Testing – Since I will be collecting and analysing high volumes of data I will be conducting load tests to make sure that R Studio or any other piece of software I will be using is performing efficiently

Unit Testing – I will use unit testing when writing code in R Studio with R and Python to make sure that all code is functional.

Conversion Testing – Will be used to make sure that any data that is converted into a new format will be uniform with any other data in the desired format and that it works seamlessly