Using Data Analysis to Understand the UK Top Music Charts

Daniel S. Pereira

Dissertation Submitted in Partial Fulfilment of the Requirement for the

Higher Diploma in Science in Data Analytics

Dublin Business School

May 2021

Declaration

I, Daniel Pereira, declare that this research is in my original work and that it has never been presented to any institution or university for the award of Degree or Diploma. Also, I have correctly referenced all literature and sources used in this work and this work is compliant with the Dublin Business School’s academic honesty policy.

Signed: Daniel Pereira

Date: May 1st, 2021

Acknowledgements

Abstract

The first line of the abstract should be flush left (not indented).

Keywords: Placeholder

Table of Contents

[1 INTRODUCTION 1](#_Toc71636086)

[1.1 Subject 1](#_Toc71636087)

[1.2 Justification 1](#_Toc71636088)

[1.3 Research Question 1](#_Toc71636089)

[1.3.1 The hypothesis to be evaluated: 1](#_Toc71636090)

[1.3.2 Conclusion: 1](#_Toc71636091)

[1.3.3 Scope and limitations of this study: 2](#_Toc71636092)

[1.4 Major contributions of this study 2](#_Toc71636093)

[1.5 Structure of the research 2](#_Toc71636094)

[2 LITERATURE REVIEW 3](#_Toc71636095)

[2.1 Introduction 3](#_Toc71636096)

[2.2 The Official UK Charts 3](#_Toc71636097)

[2.3 CRISP-DM methodology 5](#_Toc71636098)

[2.4 What data analytics techniques can be used? 7](#_Toc71636099)

[2.5 Overview of previous related work 7](#_Toc71636100)

[2.6 Conclusion 8](#_Toc71636101)

[3 RESEARCH METHODOLOGIES 8](#_Toc71636102)

[3.1 Introduction 8](#_Toc71636103)

[3.2 Business understanding 8](#_Toc71636104)

[3.2.1 Business background: 9](#_Toc71636105)

[3.2.2 Business and Data Mining Objectives: 10](#_Toc71636106)

[3.2.3 Business and Data Mining Success Criteria: 10](#_Toc71636107)

[3.3 Research methodology and strategy 10](#_Toc71636108)

[3.3.1 Project Plan: 10](#_Toc71636109)

[3.3.2 Tools and techniques: 10](#_Toc71636110)

[3.4 Conclusion 11](#_Toc71636111)

[4 DATA UNDERSTANDING 11](#_Toc71636112)

[4.1 Introduction 11](#_Toc71636113)

[4.2 Collect data 11](#_Toc71636114)

[4.3 Describe data 13](#_Toc71636115)

[4.4 Explore data 14](#_Toc71636116)

[4.5 Integrate data 15](#_Toc71636117)

[4.6 Verify Data Quality 15](#_Toc71636118)

[4.7 Conclusion 15](#_Toc71636119)

[5 DATA PREPARATION 16](#_Toc71636120)

[5.1 Introduction 16](#_Toc71636121)

[5.2 Select Data 16](#_Toc71636122)

[5.3 Clean Data 16](#_Toc71636123)

[5.4 Construct Data 16](#_Toc71636124)

[5.5 Integrate Data 16](#_Toc71636125)

[5.6 Format Data 16](#_Toc71636126)

[5.7 Conclusion 16](#_Toc71636127)

[6 MODELLING 16](#_Toc71636128)

[6.1 Introduction 16](#_Toc71636129)

[6.2 Select Modelling Technique 16](#_Toc71636130)

[6.3 Generate Test Design 16](#_Toc71636131)

[6.4 Build Model 16](#_Toc71636132)

[6.5 Assess Model 16](#_Toc71636133)

[6.6 Conclusion 16](#_Toc71636134)

[7 EVALUATION 17](#_Toc71636135)

[7.1 Introduction 17](#_Toc71636136)

[7.2 Evaluate the results 17](#_Toc71636137)

[7.3 Review process 17](#_Toc71636138)

[7.4 Determine next steps 17](#_Toc71636139)

[7.5 Conclusion 17](#_Toc71636140)

[8 DEPLOYMENT 17](#_Toc71636141)

[8.1 Introduction 17](#_Toc71636142)

[8.2 Plan deployment 17](#_Toc71636143)

[8.3 Plan monitoring 17](#_Toc71636144)

[8.4 Produce final report 17](#_Toc71636145)

[8.5 Review project 17](#_Toc71636146)

[8.6 Conclusion 17](#_Toc71636147)

[9 CONCLUSION AND RECOMMENDATIONS 17](#_Toc71636148)

[9.1 Conclusion of this research 17](#_Toc71636149)

[9.2 Recommendations 17](#_Toc71636150)

[9.3 Theoretical and Managerial Contribution 17](#_Toc71636151)

[9.4 Future Research 18](#_Toc71636152)

[9.5 Personal Reflection 18](#_Toc71636153)

[10 Bibliography 18](#_Toc71636154)

[11 APPENDICES 19](#_Toc71636155)

[11.1 Appendix A 19](#_Toc71636156)

[11.2 Appendix B 19](#_Toc71636157)

List of Figures

[Figure 1: THE FIRST RECORD HIT PARADE COMPILED BY PERCY C. DICKENS. (The New Musical Express (NME), 1952) 4](#_Toc71636158)

[Figure 2: Phases of the CRISP-DM reference model. Sources: (Chapman, The CRISP-DM User Guide, 1999) (Chapman, et al., 2000) 6](#_Toc71636159)

[Figure 3: Four level breakdown of the CRISP-DM methodology Sources: (Chapman, The CRISP-DM User Guide, 1999) (Chapman, et al., 2000) 7](#_Toc71636160)

[Figure 4 - Webpage for the Official Singles Chart for the week of 14/04/1985 / Source: OCC 12](#_Toc71636161)

[Figure 5 - Screenshot of source code of the web scrapper 13](#_Toc71636162)

List of Tables

[Table 1 - Table of list of inputs 9](#_Toc71636163)

[Table 2 - Table of attributes from dataset 14](#_Toc71636164)

[Table 3 - Top ten songs that most appeared on the charts 15](#_Toc71636165)

Using Data Analysis to Understand the UK Top Music Charts

# INTRODUCTION

## Subject

This paper explores and describes the importance of charts to measure the success and influence of a song, artist, and their music along the years. The research focuses on music charts provided by the Official Charts Company in UK. Also, aims to show the various techniques of data analysis to help students on the understanding of data gathering, data transformation and data analysis.

## Justification

If we observe our day-to-day, we will find out that music follows people in almost every moment of their life, be meaningful and happy or unpleasant moments. Songs can be easily related to love, motivational and inspirational, political or a way to express pain, suffer and sadness.

Based on the above, two factors were critical to realize this research. First, the importance of music as a product consumable by billions of people that movement an industry that grew by 8.2% and generated US$ 20.2 billion in wholesale revenues in 2019 (IFPI, 2020).

Second, the opportunity to practice the techniques for analysis of data learned along the course of Higher Diploma in Science in Data Analytics. The concept of data science and data analytics is important for any business to understand problems facing, and to explore data in meaningful ways. With the use of techniques to collect, organise, interpret, structure and present data into useful information to be used by decision-makers to obtain business gain.

## Research Question

The aim of this research is to investigate the question *what kind of information can be retrieved from the music charts?*

### The hypothesis to be evaluated:

Need text

### Conclusion:

A conclusion about the introduction and not the project.

### Scope and limitations of this study:

The scope of this research is to retrieve information from the Official UK Top 100 Music Chart. The focus is to use simple analysis techniques and learn about the various data science techniques without being restricted to any specific. There is also a limitation on what information can be obtained from the chart website and how the data is organised. Because the same music can have different versions with another featuring artists and it also can be listed on chart at different positions, and due time constraints, I have not dedicated to manually clean the full dataset and consolidate such scenarios and will consider same songs as a different entries as per available on retrieved data.

## Major contributions of this study

Need text

## Structure of the research

This research is organised into various chapters as described below:

Chapter 2 contains the literature review as given:

1. Literature theme one describes an overview of the music charts in the UK market as the “only official, trusted weekly barometer of what is popular in music now. The Official Charts Company have been a central part of British popular culture for over 60 years." (Official Charts Company, 2021)
2. Literature theme two describes the CRISP-DM methodology as being a tentative of an industry standard for data science and analytics research.
3. Literature theme three describes an overview of the data analytics techniques. It includes the importance of experiment different to avail the best suitable for the intended research.
4. Literature theme four describes an overview of the earlier research papers. This section studies the work of science of Music Retrieval Information.

Chapter 3 contains the research methodology in which I explain how I have chosen the research hypothesis, strategy, and business case overview.

Chapter 4 describes the preparation of data from collection to exploration. It includes details for programming languages and tools used and a use case development process overview.

Chapter 5 contains an understanding of the main dataset and how it was integrated with various sources.

Chapter 6 contains the creation of a model to be used in the research, the techniques to build and to asset it.

Chapter 7 describe the evaluation of the model and finds from the research. It includes a review of the process and the next steps to cover the gaps found.

Chapter 8 contains artefact design and development with a production of the final report.

Chapter 9 contains an in-depth discussion of the research. It reviews the objectives and how those were completed with the help of the adopted methodology. It also details the various limitations and proposes areas for future research.

# LITERATURE REVIEW

## Introduction

In this chapter, I would be presenting some general concepts related to the research fields of study doing a comprehensive review of literature available of the main topics such as the company behind the official music charts in the UK, the methodology for data analysis and techniques that I found useful for this research. Also, I will explore previous research for the main subject of this study and provide a conclusion on how I have planned to conduct my research.

## The Official UK Charts

According to the Official Charts Company (OCC) website[[1]](#footnote-1), “the story of the UK’s Official Charts dated from November 1952, when the then publisher of the New Music Express Percy Dickins decided that he needed some method to encourage advertisers to his recent music paper.” The first Singles Chart published in the New Music Express (NME) on Friday 14th of November 1952, is the Top 12, although it is a collection of 15 singles. As explained by (Warnik, Brown, & Kutner, 2004), despite that, before 15th of February 1969, there was not a single chart considered as official, the British Broadcasting Corporation (BBC)[[2]](#footnote-2) used aggregated results from diverse sources, the NME’s list has been reviewed and today, the OCC considers that as the official songs list. In 1969, the BBC has commissioned the British Market Research Bureau (BMRB) to compile the UK’s charts on their behalf. From that moment the UK’s official charts were born and adopted by the industry.



Figure 1: THE FIRST RECORD HIT PARADE COMPILED BY PERCY C. DICKENS. (The New Musical Express (NME), 1952)

The history and importance of the UK music charts is in many ways related to the Top of the Pops, a television programme made by the BBC and broadcasted on TV channels BBC One and later BBC Two for over 40 years. As per the description of an episode aired on 1st of April 2011, “1964 saw the birth of a very British institution. Spanning over four decades, Top of Pops has produced many classic moments in pop culture.” (BBC, 2011)

The TV show used the compiled list of top songs to present artists and performance of the number one in the charts in the end. For many people, that programme was the only way to discover new artists and know who was leading the charts.

The BBC Radio 1, which started service in 1967, has a fixed show titled “The Official Chart on Radio 1” where the top songs are revealed for the first time to listeners few hours from a compilation by OCC.

The OCC company is a joint venture operation owned by the British Phonographic Industry[[3]](#footnote-3) (BPI) and the Entertainment Retailers Association[[4]](#footnote-4) (ERA). The company manages the chart compiled by contractor Kantar[[5]](#footnote-5) which collects sales data from physical and digital audio and video releases. (Official Charts Company, 2021)

## CRISP-DM methodology

In recent years, we have seen large inventions and innovations on the study of Data Analytics, due to increase of access to the Internet, an overload of information is generated and involves the work to extract, process and manage this information which is commonly referred as “Big Data”. The success of a data analytics project depends on an assertive mix of tools and skilled people. Also, it requires an effective project management and a methodology to help to understand and guide along the complex process. As mentioned by (Jordan & Michell, 2015), due to the higher availability of computing resources process and performance, machine learning, and data processing have been able to tackle more complex tasks and mathematical and statistical problems.

The industry has been trying to establish themselves a standard process to guide the development of such complex data system (Azevedo & Santos, 2008).

This research focusses on **CR**oss-**I**ndustry **S**tandard **P**rocess for **D**ata **M**ining (CRISP-DM), a non-proprietary data mining model. It was first introduced in 1996, by a consortium composed by “veterans” of Daimler-Benz (now Daimler)[[6]](#endnote-1), SPSS (now as part of IBM)[[7]](#endnote-2) and NCR (known as Teradata)[[8]](#endnote-3).

According to (Chapman, et al., 2000), the process or methodology of CRISP-DM consists of six phases: *Business Understanding*, *Data Understanding*, *Data Preparation*, *Modelling*, *Evaluation*, and *Deployment*. The sequence of the phases is not rigid. Moving back and forth between distinct phases is always needed and it does not end once a solution is deployed because a new solution can always benefit and be trigged by the previous ones. Further in this research I will describe each one of the phases and the how I have used to support my studies.

The Figure 2 shows the phases proposed for a data mining process and the frequent dependencies between the phases represented by the arrows, while the outer circle symbolizes that one finding involves in the next one:

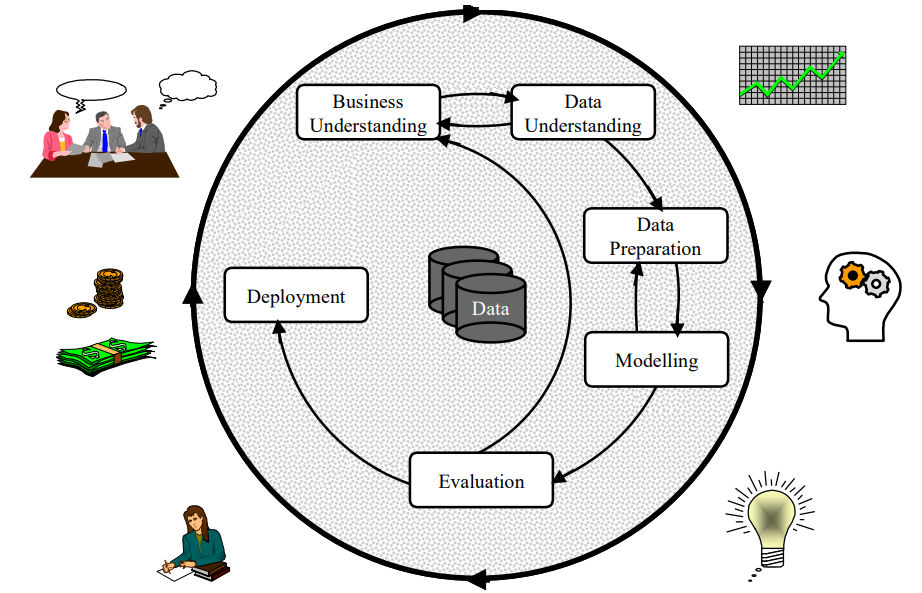


Figure 2: Phases of the CRISP-DM reference model. Sources: (Chapman, The CRISP-DM User Guide, 1999) (Chapman, et al., 2000)

The main objectives and benefits of CRISP-DM according to Peter Chapman in the User Guide presentation (Chapman, The CRISP-DM User Guide, 1999) are:

* Reduce skills required for knowledge discovery
* General purpose (i.e., stable across varying applications)
* Tool and technique independent
* Capture experience for reuse

The CRISP-DM consist of sets of tasks described at four levels of abstraction (from general to specific): phase, generic task, specialized task, and process instance.

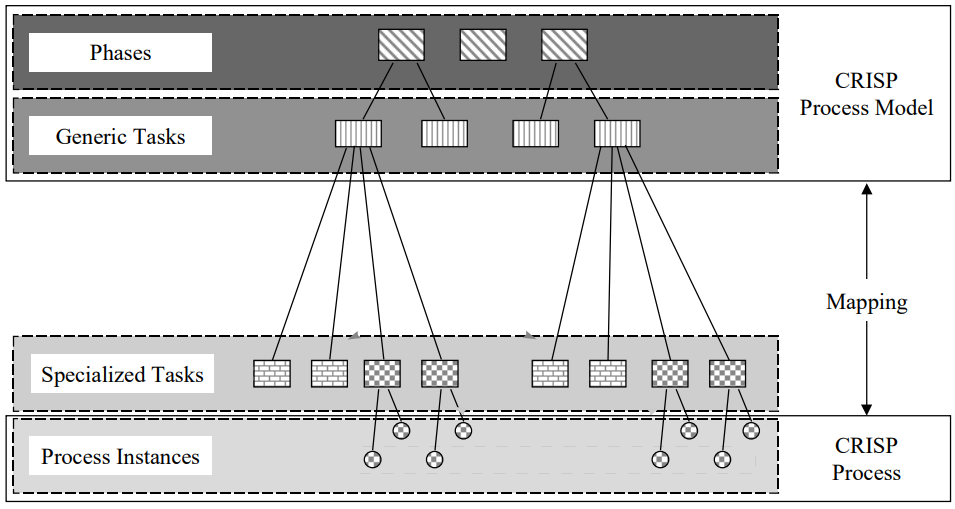


Figure 3: Four level breakdown of the CRISP-DM methodology Sources: (Chapman, The CRISP-DM User Guide, 1999) (Chapman, et al., 2000)

The Figure 3 above shows the data mining process organised into a number of phases, each phase consists of several second-level generic tasks, because it is intended to cover all possible situations but being complete and stable as possible. The third level, the specialised task level, is the place to describe ‘how’ the actions in the generic tasks should be conducted. At the fourth level, the process instance, is a record of the actions, decisions, and results of an actual data mining engagement. (Chapman, et al., 2000)

## What data analytics techniques can be used?

Need text

## Overview of previous related work

The literature is vast with research regarding the music industry, also the science of Music Retrieval Information (MRI). This research provides an overview of the official UK music charts and the kind of information could be retrieved from it.

(Bressan, 2003) studied different media types and analyse questions like “what data do we wish to extract from the music?” and “how do we intent to classify it?”. It tries to study the cultural-sociological and the more strictly computer-technical aspects.

(Dubini & Provera, 2008), addresses different innovation strategies by investigating the competitivity dynamics in the launch of new artists in the US music market. It concludes that partnership is a winning competitive formula in leading innovation into the charts.

Another topic of studies of music data analysis is “Hit Song Science” (HSS), the attempting of predict whether a given song will be a commercial success prior to its distribution, based on its audio. In (Ni, Santos-Rodrıguez, Mcvicar, & De Bie, 2011), explored the subject with a predictive model obtaining optimistic results. They had a goal of distinguishing the top five from the bottom ten in a top forty music list. In more depth, (REIMAN & ORNELL, 2018), investigated how machine learning algorithms could be suited for classify a music track as a hit or non-hit based on audio features.

## Conclusion

As we changed the ways to consume music from album to singles, and from physical media to streaming services, the music charts companies must adapt to keep relevant into this new Internet and social media world. The Official Charts Company, as the example of this study, invested in data analysis and information provider to third-party companies where you can buy valuable data regarding listening audience, market campaigns and listeners engagement.

As data itself means nothing without a given context and right tools to transform it in relevant information, I have introduced the method that I intent to use in this research. The use of CRISP-DM as a data mining method for the best knowledge of this author is still not complete. As a professional Business Analyst, I have always tried to find a standard that could be applied to all situations without many customisations and as same as many process or methods, this is impossible because the process of transform information into knowledge depends on how individuals learn. Attempt to present CRISP-DM was to focus on majority of literature available for study of data analysis techniques.

Can conclude that the music business can be avail of past information to improve itself for new audiences and can use the knowledge of academic studies of data mining and analytics to support new discoveries and its business.

# RESEARCH METHODOLOGIES

## Introduction

In this chapter, I will focus on methods of research and the business dataset that motivated me to continue exploring the possibility to complete this study.

This is the Business Understanding phase in the methodology that focusses on project goals and requirements from business needs that are later transformed into data mining goals or problem definitions. As the outcome of the phase, the inputs and constraints are identified and a plan for the whole analysis process is created.

## Business understanding

The music industry – whether an artist, label manager, A&R[[9]](#footnote-6) representative – will need to focus on what the music charts are and how they work. Because the marketing involved in such lists, it can turn an unknown artist or song into a popular mainstream success, thanks to the power of radio stations airplay, digital downloads, music streaming services and social media rankings. The audience is used to let the music charts tell them what to listen to, and it reflects on which direction the music industry is going for the that period.

Historically, music charts are lists based on popularity of songs and albums. Used as a barometer of success for entrance or veterans in the music industry.

There are so many different charts with unique calculations and factors. Mainly calculated electronically whilst also involving human elements.

### Business background:

The Official UK Top 100 music chart[[10]](#footnote-7) is compiled by the OCC, based on official sales of downloads, CD, vinyl, audio streams and video streams. While the Top 40 songs are broadcasted on BBC Radio 1 and MTV, the full Top 100 is published exclusively on *OfficialCharts.com* website.

Despite of the chart being freely available on OCC’s website, not all details of a song are, for this reason the following are a limited list of inputs:

|  |  |
| --- | --- |
| **Input** | **Description** |
| *Position this week* | current position of the single on the chart for the period |
| *Last week position* | position of the single on the chart on previous week |
| *Title* | song title |
| *Artist* | the registered artist of the single. It can be solo singer, band, group or featured artists |
| *Label* | the record company of the single |
| *Peak position* | the maximum position achieved by the single on the chart |
| *Weeks on Chart* | number of weeks the single appears on the chart |

Table - List of inputs available on OCC’s website

As the studies of music information retrieval can be complex to the point of being considered as a science itself, the following are constraints imposed to this research as limitation of scope:

1. The work focussed on the Official UK Top 100 Music Chart.
2. Despite of data is available since first list on November 14th 1952, it was only on January 2nd 1983, the list became a compilation of top 100 songs.
3. The data relies on information published on OCC’s website and is passive of human error inputs.
4. Due marketing reasons a same song can appear more than once just because it has a miss-spelling on title or introduced another artist as collaborator.

### Business and Data Mining Objectives:

The primary goals from the researcher perspectives are:

1. How does a song get into the top charts?
2. Are the charts still relevant to analyse a success of a song or artist?
3. Who are the artists who more position their song to the charts?
4. Is the genre of the music matters to make it to the charts?
5. Will a release date of a song significantly affect the changes to make it to the top of the charts?

### Business and Data Mining Success Criteria:

The factor critical of success found by this research are:

1. Deliver an easy-to-read data insight from music retrieval information
2. Demonstrate whether the number of artists has increased or decreased between decades
3. Demonstrate any significant variation on attributes of top 10 songs for a period

## Research methodology and strategy

This research is focused on the UK music industry and in special the singles charts provided by the Official Chart Company (OCC), considered the official source of music ranking and a reliable start point for music retrieval work (Official Charts Company, 2021).

The database is updated on weekly basis with information of the Top 100 Singles that made the chart for the period. As the consolidation of the top music charts were not consistent on beginning when it was created, I have access various sources to enhance the quality of the data and for those cases an explanation note will be made aware.

### Project Plan:

As following the plan to achieve the business and data mining goals:

* Text

### Tools and techniques:

To support the various methods for data mining the following are the tools and techniques applied:

* **Internet browser:** used to search for different content for this research
* **Visual Studio Code:** used for writing source code of web scrapper
* **Python 3.8:** code language used for obtaining data from the Official Chart Company’s website[[11]](#footnote-8)
* **Scrapy 2.4.1:** Python library used to create a web spider to retrieve information from the Official Chart Company’s website
* **Microsoft Excel:** used to open .CSV data files and to transform and enhance data for better analysis

## Conclusion

Why have I decided to do this research? What was the motivation?

As an introduction to this research, I have described the motivations that helped me to start a journey of studies of information that can be retrieved and analysed from songs. I have given an overview of the official music charts from the United Kingdom and the importance of itself for the pop culture market.

Given the inputs and constraints of such business, my expectation must limit to analyse potential relationship between the top singles in certain period and provide visual insights of performance on the chart.

# DATA UNDERSTANDING

## Introduction

According to (Chapman, et al., 2000), the data understanding phase starts with a first data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data or to detect interesting subsets to form hypotheses for hidden information.

## Collect data

The main data source for this research is the list of top 100 music appearing on OCC’s website since it started in 14th November 1952.

Because the only official data available is from the OCC’s website, firstly, I have contacted the company and retrieved authorization to extract information for this research under the condition to always mention the source.

As shown on the Figure 4 below, the website is composed by static pages for each weekly chart. Once loaded, it displays a list of top 100 music singles, the artist and other classification information such as the position on previous week, number of weeks on charter and the highest position ever to reach.

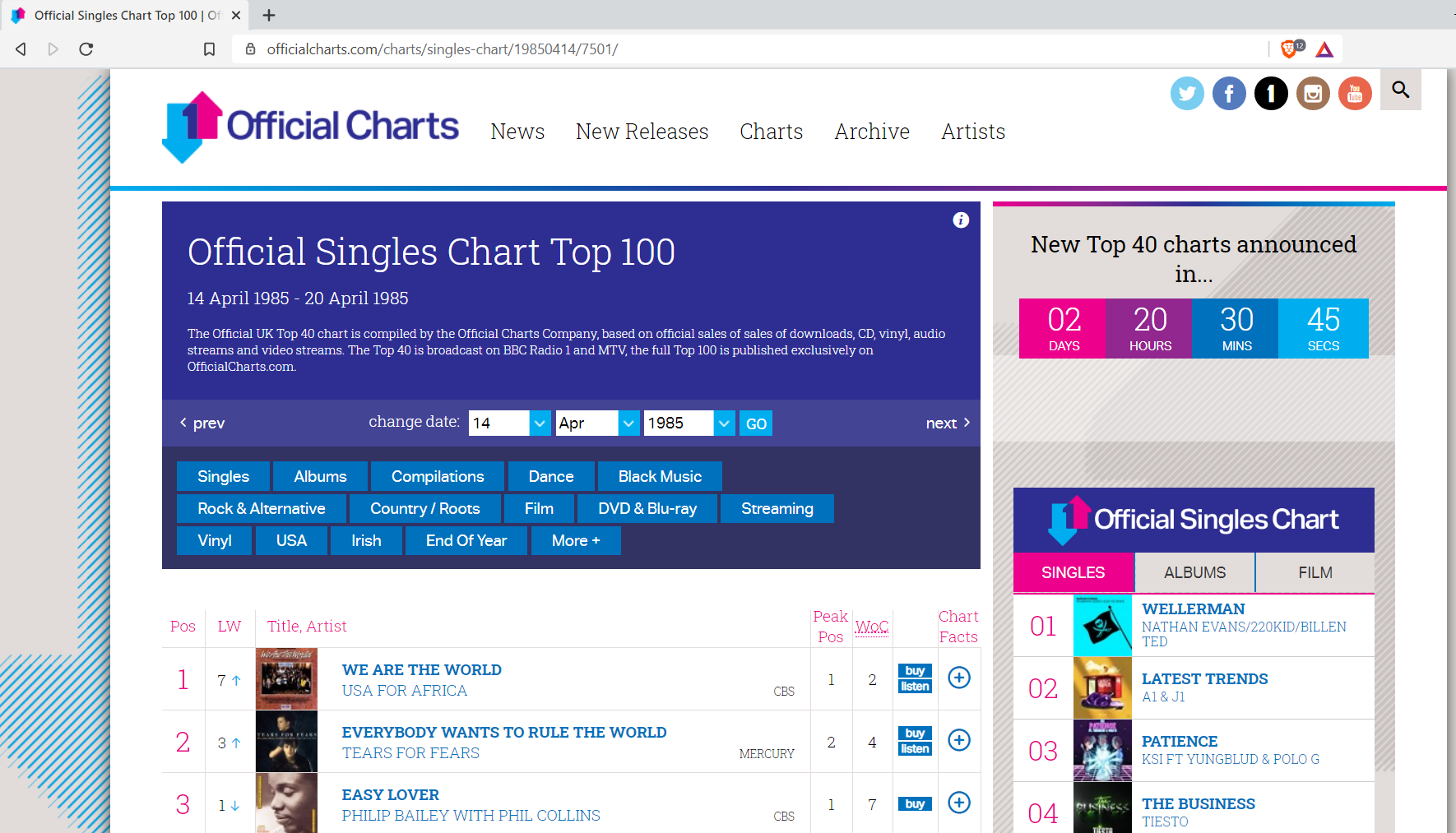


Figure 4 - Webpage for the Official Singles Chart for the week of 14/04/1985 / Source: OCC[[12]](#footnote-9)

To collect data, I have developed a ‘web spider’ routine to read each HTML tag of the page and store the information associated with it. For such work, I used samples of code in Python available by the open-sourced library Scrapy[[13]](#footnote-10). The Figure 5 below shows the code setup in Visual Code application and the full source code is available on Appendix of this research as well on my GitHub[[14]](#footnote-11).

Graphical user interface, text, application

Description automatically generated

Figure 5 - Screenshot of source code of the web scrapper

As output of the crawling process, a dataset is generated in format of a JSON file for each year of the chart. The next step to facilitate analysis, I have created a utility class to convert the JSON file into a CSV where I could use other tools such as Excel and Tableau for data manipulation and visualization.

## Describe data

The dataset is composed of 69 .csv files, following a naming convention of OfficialUKCharts1952.csv, OfficialUKCharts1953.csv, OfficialUKCharts1954.csv, and so on.

A combined Excel format file (OfficialUKCharts\_Combined\_Raw.xlsx), was created to facilitate the data manipulation. There is a total of 268,998 rows for the period of first chart on 14/11/1952 and 25/12/2020.

### Working files

To preserve the original data files (RAW files), and to keep this research as tool independent, a new work dataset file was created to manipulate data easily. This file is identified as (OfficialUKCharts\_Dataset\_Analysis.xlsx).

Firstly, to work on the dataset some attributes were added to index due the volume of rows and classify information for each song.

In a tentative to remove duplicates for a case a song appears in a sequence of weeks, it was given index values for artists with unique name, song tracks with unique name and grouped the track and artist data to give them an index value as well.

Graphical user interface, application, email

Description automatically generated

Figure - List of Artist data with index column

Table

Description automatically generated

Figure - List of Label companies with index column

It is composed of all songs a list of songs from the top 100 chart for each week. As following are the attributes.

| **Attribute** | **Type** | **Description** | **Sample** |
| --- | --- | --- | --- |
| counter\_week | Numeric | A simple incremental counter of the week extracted information. | 10 |
| chart\_week | Date | Date reference of the weekly chart | 1985-04-07 |
| chart\_day | Numeric | Day of the weekly chart | 07 |
| chart\_month | Numeric | Month of the weekly chart | 04 |
| chart\_year | Numeric | Year of the weekly chart | 1985 |
| chart\_pos | Numeric | current position of the single on the chart for the period from 1 to 100 | 100 |
| track | Text | song title | WE ARE THE WORLD |
| track\_idx | Numeric | Custom field, autogenerated a unique identifier of the song track | 9295 |
| artist | Text | the registered artist of the single. It can be solo singer, band, group or featured artists | USA FOR AFRICA |
| artist\_num | Numeric | The identification code of the artist into OCC’s database | 22419 |
| artist\_idx | Numeric | Custom field, autogenerated a unique identifier of the artist | 3416 |
| label | Text | the record company of the single | CBS |
| label\_idx | Numeric | Custom field, autogenerated a unique identifier of the label company | 35 |
| last\_week | Numeric | position of the single on the chart on previous week | 10 |
| peak\_pos | Numeric | the maximum position achieved by the single on the chart | 10 |
| weeks\_on\_chart | Numeric | number of weeks the single appears on the chart | 10 |

Table - Attributes from consolidated dataset

The dataset suffers of the duplication problem on that artists have non-unique name will appear in different instances of the charts in a non-correlated week. Also, the same song is released under different version featuring other artists and making a different record accountable. As far this researcher is aware, there is no mechanism to automatically check the duplicates. The duplicates may impose a problem on interpretation of data but for this research it will be considered as it appears on the charts.

## Explore data

During the analysis of the charts, I could conclude that is not possible to define unique songs since the same song can have different version recorded by different artists. In this research, unique songs is considered unique record found with the combination of track and artist.

Analysing the data attributes in detail, the first questions to exam are:

1. How many unique number one songs exist between 1952 and 2020?

Based on above, 43180 entries was found.

1. Top five genres for the period? All time and per decade
2. Top ten songs that most appeared on the charts

The songs that most appeared on charts means the number of times a track and artist appeared on charts. To find the top rank was just order by total count from higher to lower.

| **rank** | **track** | **artist** | **chart\_count** |
| --- | --- | --- | --- |
| 1 | MR BRIGHTSIDE | KILLERS | 240 |
| 2 | CHASING CARS | SNOW PATROL | 166 |
| 3 | PERFECT | ED SHEERAN | 140 |
| 4 | MY WAY | FRANK SINATRA | 133 |
| 5 | SEX ON FIRE | KINGS OF LEON | 124 |
| 6 | THINKING OUT LOUD | ED SHEERAN | 118 |
| 7 | SHOTGUN | GEORGE EZRA | 118 |
| 8 | CHANDELIER | SIA | 114 |
| 9 | WHATEVER | OASIS | 112 |
| 10 | I GOTTA FEELING | BLACK EYED PEAS | 109 |

Table - Top ten songs that most appeared on the charts

1. Top five longest number one songs
2. Female artists with the most top 10 songs on the charts
3. Male artists with the most top 10 songs on the charts
4. Bands with the most top 10 songs on the charts
5. Duet with the most top 10 songs on the charts
6. Trio with the most top 10 songs on the charts
7. Groups with the most of top 10 songs on the charts
8. Artists on bottom of the charts who never ranked more than position 90?
9. Artists with ‘one hit wonder’ and have only one song as top 10 on the charts?
10. Top 5 artists by season: sprint, summer, autumn and winter
11. Most popular label with most number one songs

## Integrate data

## Verify Data Quality

## Conclusion

# DATA PREPARATION

## Introduction

The data preparation covers all activities needed to construct the final dataset. (Chapman, et al., 2000). Tasks include, select, clean, construct and integrate data for modelling tools.

## Select Data

## Clean Data

## Construct Data

## Integrate Data

## Format Data

## Conclusion

# MODELLING

## Introduction

## Select Modelling Technique

## Generate Test Design

## Build Model

## Assess Model

## Conclusion

# EVALUATION

## Introduction

## Evaluate the results

## Review process

## Determine next steps

## Conclusion

# DEPLOYMENT

## Introduction

## Plan deployment

## Plan monitoring

## Produce final report

## Review project

## Conclusion

# CONCLUSION AND RECOMMENDATIONS

## Conclusion of this research

In this research, I used the available data from the UK Official Charts of 100 songs, which included over 263,000 records, and over 43,000 songs releases over a period of (1952 and 2020). I studied the various attributes of the music chart as well explored the evolution of a song and an artist along the years.

Despite of this research be limited to the dataset that I have defined, it still has a theoretical value.

## Recommendations

## Theoretical and Managerial Contribution

## Future Research

## Personal Reflection

# Bibliography

Azevedo, A. I., & Santos, M. F. (2008). *KDD, SEMMA and CRISP-DM: a parallel overview.* Instituto Politécnico do Porto. Instituto Superior de Contabilidade e Administração do Porto. Porto: Instituto Politécnico do Porto. Instituto Superior de Contabilidade e Administração do Porto. Retrieved April 20, 2021, from http://hdl.handle.net/10400.22/136

BBC (Director). (2011). *Top of the Pops 1964 to 1975 - Big Hits* [Motion Picture]. The United Kingdom. Retrieved April 20, 2021, from https://www.bbc.co.uk/iplayer/episode/b00zwrn5/top-of-the-pops-1964-to-1975-big-hits

Bressan, F. (2003). *An overview of music information retrieval (MIR) databases.* Udine: Universita Degli Studi Di Udine.

Chapman, P. (1999). The CRISP-DM User Guide. *Brussels SIG Meeting* (p. 14). Brussels: NCR Systems Engineering Copenhagen. Retrieved April 20, 2021, from http://lyle.smu.edu/~mhd/8331f03/crisp.pdf

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0: Step-by-step data mining guide.* The CRISP-DM Consortium. Retrieved April 20, 2021, from ftp://ftp.software.ibm.com/software/analytics/spss/support/Modeler/Documentation/14/UserManual/CRISP-DM.pdf

Dailmer AG. (2021, April 20). *1995-2007. "Word Corp." vision*. Retrieved from Daimler AG web site: https://www.daimler.com/company/tradition/company-history/1995-2007.html

Dubini, P., & Provera, B. (2008). Chart Success and Innovation in the Music Industry: Does Organizational Form Matter? *Journal of Media Business Studies*, 41-62. doi:https://doi.org/10.1080/16522354.2008.11073460

IBM. (2009, July 28). *IBM to Acquire SPSS Inc. to Provide Clients Predictive Analytics Capabilities*. Retrieved April 20, 2021, from IBM New Releases: https://www-03.ibm.com/press/us/en/pressrelease/27936.wss

IFPI. (2020). *Global Music Report 2020 - The Industry in 2019.* London: IFPI. Retrieved April 20, 2021, from https://www.ifpi.org/resources/

Jordan, M. I., & Michell, T. (2015, July 17). Machine learning: Trends, perspectives, and prospects. *Science, 349*(6245), pp. 255-260. doi:https://doi.org/10.1126/science.aaa8415

Ni, Y., Santos-Rodrıguez, R., Mcvicar, M., & De Bie, T. (2011). *Hit Song Science Once Again a Science?*

Official Charts Company. (2021, April 20). *Who we are - The Official Charts*. Retrieved April 20, 2021, from Official Charts.com: https://www.officialcharts.com/who-we-are/the-official-charts/

REIMAN, M., & ORNELL, P. (2018). *Predicting Hit Songs with Machine Learning.* STOCKHOLM: EXAMENSARBETE INOM TEKNIK.

Teradata. (2020, August 20). *Our History*. Retrieved April 20, 2021, from Teradata: https://www.teradata.com.au/About-Us/Our-History

The New Musical Express (NME). (1952, November 14). Announcing the first Record Hit Parade. (R. Sonin, Ed.) *The New Musical Express*(305), p. 8.

Warnik, N., Brown, T., & Kutner, J. (2004). *The Complete Book of the British Charts: Singles & Albums* (3rd ed.). Omnibus.

# APPENDICES

## Appendix A

## Appendix B

Have you imagined the season or weather of a region defining what kind of music people will hear on the radio? In the summer we are all used to hear happy songs, energetic and animated.

Created in 1920 e until today the media channel that reaches people in all parts of the world, the radio continues being one of the main tools to discover what song is a hit or not. But, not always this is something defined by the audience of listeners.

Excluding those dictatorial regimes where the goal is to overtake the liberty of expression and basic civil rights, and not to control or censorship a specific song, there is still censorship in music, and it is not exclusive. As an example, pop artists such as Madonna, Janet Jackson and Katy Perry by different and specific reasons already have and continue to be experiencing a kind of boycott from radios and some reasons are quite hilarious when we get to know the motivation behind some songs being pulled off from radio streams.

How music is evolving?

What makes a song a hit?

Why do you listen to a song more often than others?

What can I analyse:

• Genre

• Lyrics

• Rhythms

What is the impact of music on life facts?

How can a life event impact on music?

How can music translate a historical fact?

How did the music changed in the past 60 years?

Why music is important?

Does it influence on our manner of thinking?

1. Official Charts Company’s website: [www.officialcharts.com](http://www.officialcharts.com) – accessed on 20 August 2020 [↑](#footnote-ref-1)
2. BBC’s website: [www.bbc.co.uk](http://www.bbc.co.uk) – accessed on 20 August 2020 [↑](#footnote-ref-2)
3. BPI’s website: [www.bpi.co.uk](http://www.bpi.co.uk) – accessed on 20 April 2021 [↑](#footnote-ref-3)
4. ERA’s website: [www.eraltd.org](http://www.eraltd.org) – accessed on 20 April 2021 [↑](#footnote-ref-4)
5. Kantar’s website: [www.kantar.com](http://www.kantar.com) – accessed on 20 April 2021 [↑](#footnote-ref-5)
6. Daimler-Benz is a car manufacture well known by the Mercedes-Benz brand. In 1998, the company has merged with the American based Chrysler and formed DaimlerChrysler AG. In 2007, after the completion of sale of Chrysler, it was approved a change of name to Daimler AG (Dailmer AG, 2021). [↑](#endnote-ref-1)
7. SPSS was an American software company formed in 1968, which provided a software of same name for statistical analysis and social science. It was acquired by IBM in 2009, since that the company was dismissed, and the product rebranded to IBM SPSS (IBM, 2009). [↑](#endnote-ref-2)
8. Teradata is an American software company formed in 1979, it was acquired by NCR (National Cash Register) a consulting service in December 1991 (Teradata, 2020). [↑](#endnote-ref-3)
9. Artist and Repertoire (A&R) – is person or group entitled for finding promising new artists for a record label or music publisher to signing it. [↑](#footnote-ref-6)
10. <https://www.officialcharts.com/charts/singles-chart/> [↑](#footnote-ref-7)
11. <https://www.officialcharts.com/charts/singles-chart/> [↑](#footnote-ref-8)
12. <https://www.officialcharts.com/charts/singles-chart/19850414/7501/> [↑](#footnote-ref-9)
13. Scrapy is an open source and collaborative framework for extracting the data from websites. It is maintained by Zyte and many other contributors. <https://scrapy.org/> [↑](#footnote-ref-10)
14. Github repository [danielpereiradotcom/DBS: Collection of projects for Data Analytics course at Dublin Business School (github.com)](https://github.com/danielpereiradotcom/DBS) [↑](#footnote-ref-11)