With this project my aim is to improve my data analysis and data science skills and to create a public repository that I can use to showcase my capabilities.

For this first project I’ve chosen a topic that I’m not familiar with and that I hope it will give me the opportunity to try different data analysis tools that I haven’t been able to try on my date to date job yet.

As a little bit of background I am a manufacturing engineer with extensive experience in the aerospace sector and in shop floor environments.

Coming back into the project, I have segregated the data collection part into a data\_creation module for tidiness purposes.

I check whether the data has already been collected (which would be stored in parquet file on the current working directory) and if not I collect the data from the following webpage using pandas:

<http://www.umdmusic.com/default.asp?Lang=English&Chart=D&ChDay=1&ChMonth=2&ChYear=1999&ChBand=&ChSong>=

I decided to collect data using parquet files from 1970s to 2020s to be able to see if there is any change in the behaviour of the data across time. I chose to scrape the data 4 times for every month in that year range.

The reason on one hand is to create enough granularity of the data, since the charts are on a weekly basis, I need to collect the data weekly. On the other hand I want to create a bigger dataset than the ones I’m used to work with.

My initial objective is to find out the characteristics that differentiate the top songs (the ones that reach the number 1 position in the charts) from the rest of the songs.

With that data I try to answer the following 3 questions:

* How long do n1 songs last in charts vs the rest of the songs?
* How many positions do n1 songs climb vs rest of the songs?
* What the starting positions of n1 songs are in charts vs the rest of the songs?

To be able to compare the n1 songs population with the rest of the population I differentiate the Artist and Title of each song that has ever reached a number one position, this is, where the this week position has ever been 1.

I then create the attributes required to answer the questions stated above:

The maximum date in the charts