The data consists of 317 songs (all in the year 2000) and their evolution in the Billboard charts. For each song the data includes the artist, the length of the song, the genre, the date it entered the charts and the date of the peak in the charts. After this, each song has a timeline of their evolution in the charts. This timeline goes on for 76 weeks, which means that songs that have not been in the charts for 76 weeks has missing data in one of these columns. An interesting next step would be to find out for how many songs this is the case.

It is reassuring to see that the count does not give missing values for any of the track-specific information (such as artist, song length or genre) or the week of entrance in the data charts. However, what can also be seen from the data is that the number of columns is larger than it should be. No song in our sample has been in the charts for longer than 65 weeks, which is why we can reduce the number of columns.

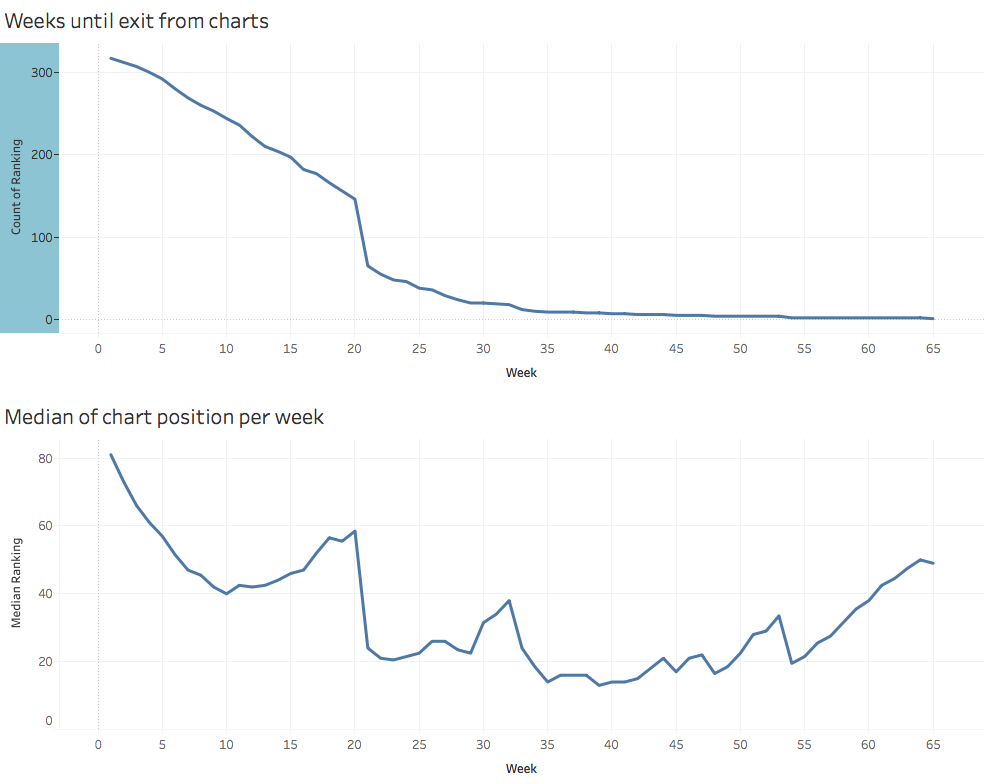
A new dataframe is created out of the previous dataframe with raw data, where columns with only missing data have been trimmed. To make handling of the data easier, the column titles for the weekly timeline are changed to just the week number in the timeline. Artist.inverted is renamed to just artist, and the first column of the data (the year 2000) has been removed.

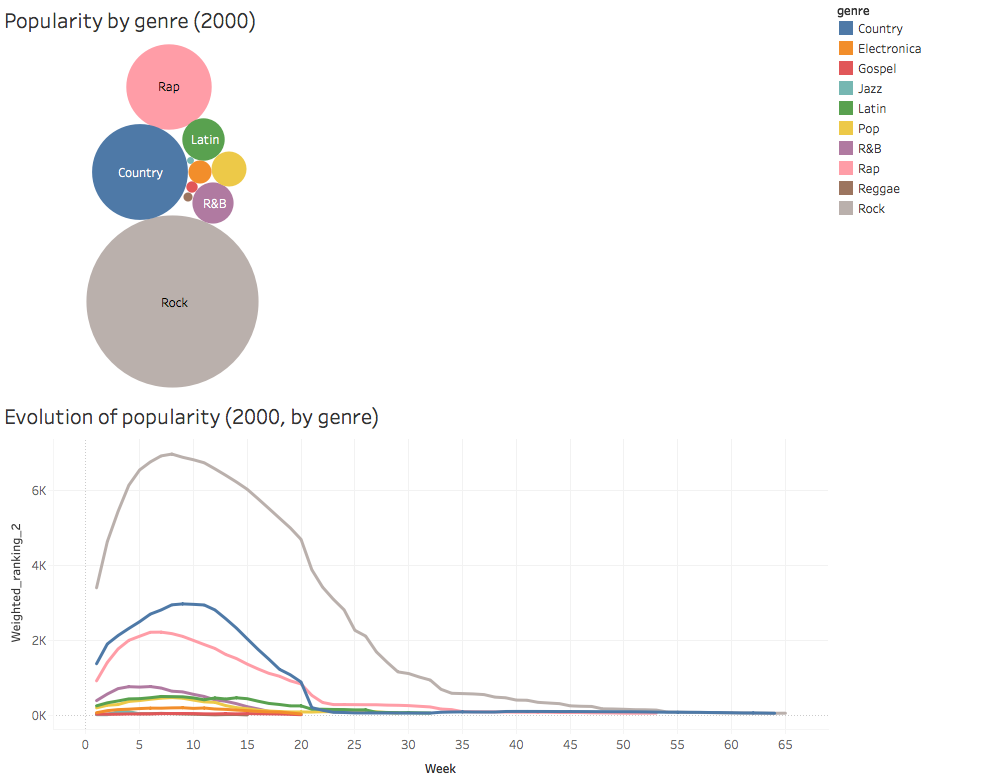
All other missing data is not yet replaced. Observations with missing data should not be removed since it indicates that the song is no longer in the charts at a specific point, and intuitively it would not make sense to replace this data with a string or an integer.

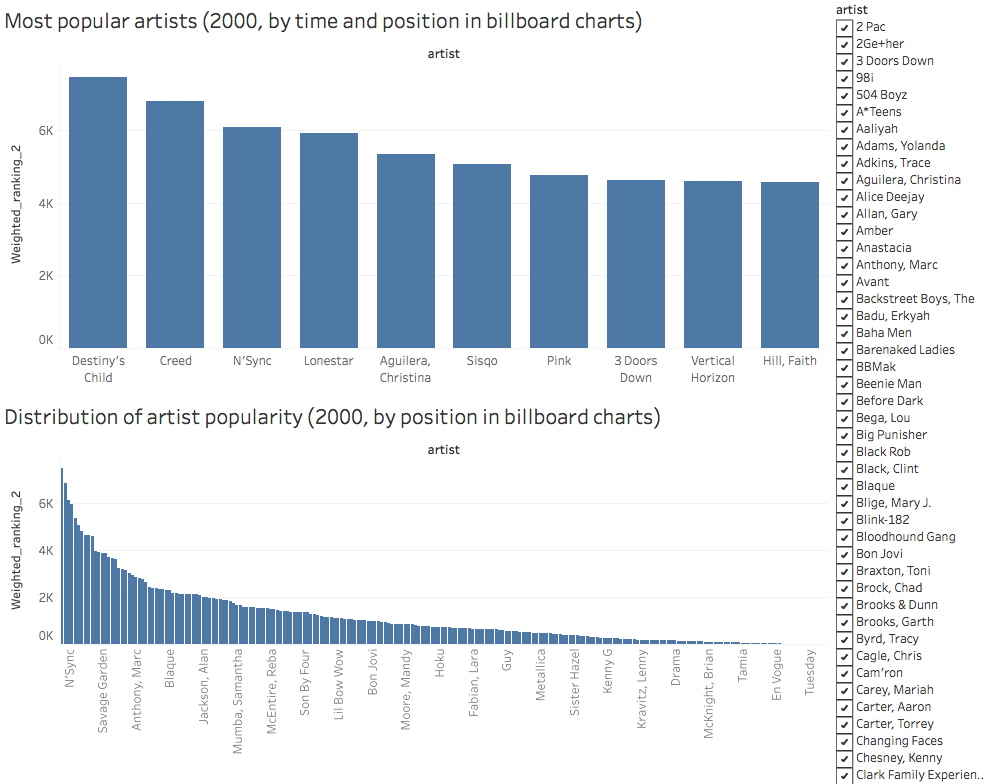
**For visualizations in Tableau, a new variable (weighted\_ranking) was created, defined as:¶**

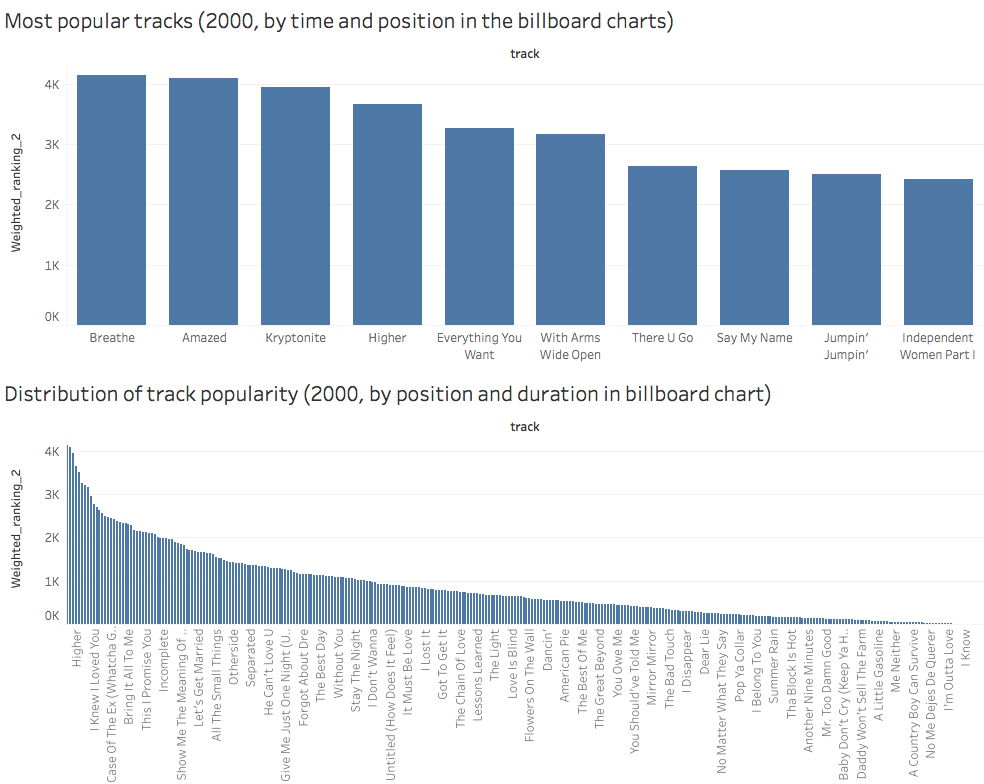
***weighted\_ranking = 100 - Ranking¶***

This assigns more weight to tracks that have been higher in the chart, and is thus a better proxy for popularity









Now that we have an overview of the popularity of each song and artist in the year 2000, it would be interesting to see how this correlates to popularity in more recent times. This could determine what type of songs are just popular at their release, and what kind of songs remain popular over a longer period of time. Youtube views (www.youtube.com) or spotify plays could provide an indication of the recent popularity of each song in our database of 2000.

* Collect data from youtube (views per song) and spotify for each song in our database.
* Append this data to our database as a new column, with a value for each song.
* Calculate a new ratio, recent\_popularity, where youtube views is divided by each song's relative popularity in the year 2000.
* Take the average of this new ratio recent\_popularity over each genre, and find out if they differ significantly per genre.

A visualization of this indicates in which genre songs are popular over a longer period of time.