Analítica de los charts Billboard US Top100 desde 1960

Al principio quería estudiar más sobre música europea, es decir, ver qué tipo de éxito tenían las canciones francesas en España o al revés, o estudiar como las canciones en español (o latino) tenían cada día más éxito. Pero o los datos no eran muy bien estructurados, o no se pudio encontrar fácilmente la nacionalidad o el idioma de las canciones, así que todo eso era un proyecto un poco largo para este primer trabajo.

Pero encontré un dataset muy bien construido de todos los charts americano (billboard Top100) desde 1958, es decir el ranking semanal de las 100 primeras canciones (según índices que obviamente han cambiado a lo largo del tiempo) desde 1958. Eso me permitía de tener una idea de las tendencias sobre una larga periodo de tiempo. El dataset era infortunadamente un poco limitado a nivel de información, así que decidí añadirle el dataset de Spotify sobre 160000 canciones desde 1920, que incluye muchas características de las canciones, como el tempo, la duración y la “danceability”. Esos no era mi foco de estudio, pero el dataset de Spotify me permitía de ver un poco las tendencias sobre todas las canciones, y no solo las canciones que entraron al menos una vez dentro del Top100.

Decidí entonces averiguar si el concepto de Long Tail, exprimido por primera vez por Chris Anderson en 2004, seguía vivo y como se encontraba. Este concepto explique, que, con Internet, los mercados de nichos agregados representaban quizá más dinero que el mercado de masas. Con un ranking de canciones, quería focalizar mi estudio sobre los No1. A lo largo de los anos, como una canción lograba a ser No1, como sobrevivía dentro del chart, cuantos artistas lograban a tener canciones No1 etc.

El trabajo del EDA tal cual no ha sido muy complicado una vez que tenía claro lo que buscaba (el notebook tal cual incluye todas mis búsquedas así que quizá no se ve claramente lo que buscaba, porque era interesado a ver otros aspectos del dataset, más allá de solamente las canciones No1). Tenía que limpiar el dataset con solamente las décadas que quería (quitar los charts de los años 50 que quedaban). El trabajo más duro ha sido para crear un Performer ID porque infortunadamente las canciones pueden ser cantadas por varios intérpretes y quería tener claro el trabajo de cada artista, que sea para canciones donde cantan solos hasta canciones donde aparecen como Featuring. Decidí separar lo que interpretaba como ‘Duet’, es decir una canción cantada por dos intérpretes al mismo nivel, de los Featurings, donde un artista está apoyado por otros artistas. Sobre este tema, es muy interesante ver como los Featurings son sobre utilizados con un cuasi 25% de las canciones en la década de 2010. Y hay unos artistas que cuasi solo trabajan con Featurings (el español Enrique Iglesias por ejemplo…). Es lo mismo que el co-branding visto en otras áreas.

El resto, la verdad, sale bastante claro del notebook y de la presentación ppt. Lo único que me surgió de este trabajo de EDA era que hay que tener cuidado con un trabajo de datos porque de los datos, siempre se puede sacar conclusiones distintas, según lo que cada uno busca. La metodología es importante.