**PROJECT REPORT**

**Extract: your original data sources and how the data was formatted (CSV, JSON, pgAdmin 4, etc).**

All of our data was in CSV format. Our files were the [most followed on Instagram](https://data.world/socialmediadata/most-followed-on-instagram) and [billboard hot 100](https://data.world/kcmillersean/billboard-hot-100-1958-2017) data sets from Data.World.

**Transform: what data cleaning or transformation was required.**

From the most\_followed.csv file, we decided to only keep the Name, Categories 1, Categories 2 and Followers. We renamed ‘Categories 1’ to ‘Categories’ and “Categories 2’ to ‘Subcategories’. Our follower counts were followed by the unit “Mæ(=)” so we used slicing to change the unit to “M” for millions. Later on, we realised we needed to remove the “M” unit all together as we needed to calculate the mean and this made it a varchar.

To clean and transform the hot\_100.csv, we first decided to keep the Performer, Song, Track Explicit and Track popularity. We then renamed the related columns to Explicit Material and Track Popularity. We noticed a NaN value in row. We discovered we may have deleted this manually when viewing the files in excel. We added "Lady Gaga & Bradley Cooper" back into this cell. We then merged the clean data frames from both CSV files on the performer’s column.

From the merged data frame, we decided to do some further cleaning and keep only the performer, followers and track popularity columns to compare the track popularity against Instagram influence. We calculated the mean of the track popularity from all songs grouped by artist. We chose to use the max function to order by the follower count as this was static per artist. This was displayed in a data frame against followers and ordered by descending track popularity.

**Load: the final database, tables/collections, and why this was chosen.**

The final database is a a comparison between Instagram Influence and Track Popularity for musicians on the top 100 charts. Of all the data we extracted, this demonstrated the relationship most clearly. Interestingly, we can see that the most popular artist is Niall Horan, though his instagram following is no remotely as high as some of the other performers. Though Selena Gomez easily has the highest Instagram following, her track popularity sits at number four overall. In conclusion, we can determine that instagram's influence on artist track popularity is not as evident as initially presumed