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 main 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. We then manipulated the data to match up with the [annual incomes of highest paid celebrities](https://www.statista.com/statistics/262925/annual-incomes-of-highest-paid-celebrities/), [most followers worldwide](https://www.statista.com/statistics/421169/most-followers-instagram/) and [top artists based on spotify monthly listers](https://www.statista.com/statistics/1032826/spotify-artists-monthly-listeners-worldwide/) data sets from Statista. pgAdmin 4 was used for this manipulation.

**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.

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.

From the merged data frame, we decided to do some further cleaning and keep only the performer, subcategory, followers and track popularity columns.

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