**ETL Project Report:**

**Project Title**: Comic Universe Transformed

**Team Members:** Brant, Karine, Johnathan, Sequoyah, Matthew

**Project Description/Outline:**

Extract csv files to be used to transform data by cleaning, summarizing, joining, and filtering to be used to load to a relational database PG Admin4

**Data sets to be information:**

* The sources of data that you will extract from. We will be extracting CSV files from websites such as: dataworld, datahub, and fivethirtyeight
* The type of transformation needed for this data (cleaning, joining, filtering, aggregating, etc).CSV files
* The type of final production database to load the data into (relational or non-relational).PG Admin4
* The final tables or collections that will be used in the production database.TBD
* **E**xtract: your original data sources and how the data was formatted (CSV, JSON, pgAdmin 4, etc).
* **T**ransform: what data cleaning or transformation was required.
* **L**oad: the final database, tables/collections, and why this was chosen.

The data comes from Marvel Wikia and DC Wikia. Characters were scraped from FiveThiryEight. Appearance counts were scraped from FiveThiryEight. The data used was formatted in CSV and available JSON. Our group decided to strictly work with CSV data. Pandas and Python was then used to transform and read the data set. The data is split into two files, for DC and Marvel, respectively: dc-wikia-data.csv and marvel-wikia-data.csv. CSV datasets then were stored into a data frame via file path location.

Transformations included cleaning data by setting different column names and setting data types to corresponding data types like int64 to be readable with pandas, creating a final data frame for both csv files.

Both final\_marvel.to\_sql(name='final\_marvel', con=engine, if\_exists='append', index=False) and final\_dc.to\_sql(name='final\_dc', con=engine, if\_exists='append', index=False) was successfully loaded into PGAdmin4 for further analysis and data manuilptation. Generic Tables 1 and 2 were created to view data and represent superhero data for data comparison.