**ETL Project Report**

**Extract:**

The original datasets were both in a csv format. The first was provided under an MIT license and was pulled from Kaggle. The second was pulled from Caltech’s open NASA Exoplanet Archive.

**Transform:**

The first data set was already well formatted and only required the column for the planet names to be changed for the merging of the data. The second dataset originally consisted of 356 columns which had to be narrowed down to the columns with the data of most immediate use. Those columns then had to be renamed to make them more understandable. Fortunately, the data came with a dictionary of terms. Lastly, the two datasets were merged on the PlanetName columns to form one dataframe. That dataframe was then converted into a list of dictionaries. Those dictionaries were then moved to a new list, dropping the keys that had no values to ensure a clean set of data for upload to MongoDB

**Load:**

A database named ETL and a collection within named extrasolar were created in MongoDB. The data was then loaded into the collection. I chose to use MongoDB over SQL because most of the planets do not have data on them in every single column. Due to the non-relational nature of MongoDB, it seemed more logical to use over the relational structure of SQL.