**ETL-PROJECT REPORT**

Team Members:

Jose Luis Bracho

Mariana Chavez

Dwight Mauk

**OBJECTIVE**

To evaluate budget, gross income and IMDb score from a list of Oscars awarded movies.

**EXTRACTION**

We pulled data from **data.world** and **Kaggle** to obtain general information on Movies and the Oscars Awards. This data was stored in csv files that we then read into data frames using Jupyter Notebook.

**TRANSFORMATION**

* Pandas

After analyzing the data, we had to clean it up. With the Movies data we were looking to find the movies between 1995 and 2015. We focused on obtaining the following information about the movie: film name, year, director, budget, gross and IMDb score. To accomplish this, we selected the columns with the information wanted, rename the columns, filter the information to dates between 1995-2015, and formatted the data. We named our index “ID” which we then used as primary key of this table.

Similarly, to Movies data, we were looking to find the Oscars winners between 1995 and 2015. We focused on obtaining the following information about the Oscars: film name, year, award and name (of the awarded). To accomplish this, we selected the columns with the information wanted, rename the columns and filter the information to dates between 1995-2015. We named our index “ID” which we then used as primary key of this table too.

* SQL/Postgres

After our data was loaded into postgres database, our table Oscars was running without issues. However, we realized table Movies was not loading the data. Upon evaluation, we realized the name if the films had extra characters that were invisible but were able to bypass this using the function “trim”. Then we successfully were able to inner join these tables.

**LOADING**

Before starting this process, we obtained the data types of each column in order to create the schema table in pgAdmin. Then we created 2 tables called movies and Oscars with same columns to match our tables in Jupyter notebook. We created a connection between pandas and pgAdmin. We loaded the 2 tables into the database in postgres.