**Final Report:**

**ETL project – United States Museums Data**

**Members: David Monstacero, Devin Sherwood, Wilson Chira**

Extract: Our original data files.

* [**Resources/museum-universe-data-file-fy-2015-q3-et8i-mnha.csv**](Resources/museum-universe-data-file-fy-2015-q3-et8i-mnha.csv) **from Data.World**
* [**Resources/museums\_separated.csv**](Resources/museums_separated.csv) **from Kaggle**
* [**Resources/museums.csv**](Resources/museums.csv) **from Kaggle**

What we are doing with these original data files is extracting the information of what we are wanting to know. We are CSV to dataframes With so much information available to us we decided to focus on the following data.

**Museums csv**

* **Museum name**
* **Legal name**
* **Museum type**
* **Street address**
* **City**
* **State**
* **Zip code**

**Museum Universe 2015 csv**

* **Common name**
* **Legal name**
* **Phone**
* **Web URL**
* **Revenue 2015**
* **Long**
* **Lat**

**Museums 2013 Separated csv**

* Museum name
* Legal name
* Revenue 2013
* Region code

Transform: Data Cleaning

Our data was pretty clean to begin with, but we did need to do some initial cleaning and transforming. We pulled the data from Museums 2013 columns “Museum name, Legal name, Annual Revenue and region code”. We then pulled the data from Museum Universe 2015 columns “Common name, Legal name, phone, web URL, revenue, longitude and latitude”. For our third file Museums we pulled “Museum name, Legal name, Museum type, Street address, City, State, and Zip code”. After pulling the columns and creating each data frame we dropped duplicate columns and cleaned the data. We sorted the legal names by alphabetical order and set the index. We then created a filtered data frame and dropped the duplicated rows within the new filtered data frame. After cleaning, dropping, and combining, we then set up the connection for the postgres data base.

Within Postgres we –