ETL Project Report

The group decided to extract, transform, and load information based on data science job positing in three cities. We focused on Los Angeles, San Francisco, and New York. The data source were all in CSV format.

We decided to use python to transform the data. The data source have five data sets of information and that includes: position, company, description, numbers of reviews, and location. The group wanted to find companies that range based off the number of reviews the company has. The number of reviews faced several issues. First, the number of reviews data set has “reviews” with the numeric value that we want to describe. Second, the data source has NaN or null amounts if the company does not have reviews.

We decided to split the number of reviews from the numeric value column by creating a new column for number of reviews only with the numeric value. We followed this by filling in the null or NaN with 0 values. We then decided to group the data set by the number of reviews for the company that is posting the job opening. We created our groups: “A” has 0-2 reviews, “B” has 3-58 reviews, “C” has 59-705 reviews, and “D” has +706 reviews. This grouping helps the person looking for a position in their three selected cities to find companies with the most reviews. In addition we cleaned up the location data set to only include the city and not the zip code. We merge the three data sets from different locations by using append to merge all the files into one.

The group used import from pandas to PGadmin to put the new database into SQL. This allows us to view top reviewed companies from 3 different locations.