ETL Project

For this ETL project we wanted to collect data to begin analyze gender pay gaps for different types of occupations. The data we collected for the ETL project were:

* A census data file in CSV format from the Enigma Public website.
* A CSV from the Federal Bureau of Statistics.
* Html table data we scrapped from PayScale’s website regarding gender pay gaps.

During the exploratory data analysis phase of the project we found and transformed non-ascii characters such as ‘\*\*’,’—‘,’(X)’ and removed unnecessary columns. In addition, we transformed and cleaned up the html table data before writing the data into a csv file in preparation for uploading to our database. After the data was cleaned, we loaded the data frames via python scripts into a MongoDB. We chose mongo DB since it will provide high performance, high availability, and automatic scaling. Finally we made a Flask app to query the database from the user.