**PROJECT SUMMARY**

Project Proposal:- (Group 6 )  
The objective of this ETL project  is  to extract two unrelated data set from [www.kaggle.com](http://www.kaggle.com/)  for selected colleges in the US states; while focusing on the “tuition”, “projected salaries after college”, and  any other “values records”, we plan to transform data by removing irrelevant, inaccurate, and duplicate  records, validate data for accuracy,  present  the data in preferred grouping and merging the two data sets into a single data. The merged data will be loaded into a SQL database.

Extraction - Brandon

Data source – www.kaggle.com  
<https://www.kaggle.com/wsj/college-salaries#salaries-by-college-type.csv>  
<https://www.kaggle.com/theriley106/university-statistics#schoolInfo.json>Transform

Extraction Process:Once we found the data needed, the files were downlanded as CSV and JSON file extentions. The CSV file was suffficient as downloaded while Panda was used extract the required data fields from Jason.

Transform - Timothy Ayoade

Panda was used to transform and merge the two data set.

Transformation Process:-The bulk of the work with the data transformation was data cleansing (see llist of cleansing carried on the two data sets). Once the two data were cleansed and relatable, they were merged with left join with the school name as the unique identifier for both data set. The merged file was reviewed, cleansed and exported as CSV for loading to SQL.

Data cleansing steps:

* Remove incomplete and inaccurate records
* Delete empty records
* Extracted only columns required
* Removed leading and lagging spaces
* Removed unwanted characters
* Change data type to enable groupby and mean

Other data efforts:

* Data aggregation – Groupby
* Data merging – left job
* Data output – Saved as csv

Load - Mohamed Abdi  
Load (SQL)

Mohamed-L: Took the transformed data and loaded it into a Postgres database. Ran into some issues with permissions but took care of it with the chmod a+rX command in the mac terminal to upload it into Postgres. We chose a relational database because as the tables get more complicated it makes it easier to make adjustments and do more complex analysis.