**ETL Project Report**

**Introduction:**

Our ETL Project consists of data of all the colleges/schools in the United States with respect to school name, rank, whether its private or public, the accreditor of the school, location of the school, website, minimum SAT score required for the intake, acceptance rate, number of FAFSA applications, average cost of the entire course i.e. tuition and fees etc. roughly we managed to found all the above information for 7200 colleges/schools. The data we gathered was from 2 different sources and the URL for them are included in this report.

**Extraction:**

For the extraction of the data we started with the hunt for data over the internet and obtained it from this API: <https://collegescorecard.ed.gov/data/documentation/> , the data was available in the JSON format and we extracted it using the API key and loaded all the data in the Jupyter notebook and then made CSV files of the data we obtained. The other data which we obtained was from this URL which was the rankings of the colleges/schools in United States as of 2017 <https://data.world/education/university-rankings-2017> , this data was available in the CSV format.

**Transform:**

For the transformation part we decided to extract the data in the form of four groups which we eventually converted it into five tables later, during this process we imported all the CSV files into the Jupyter notebook and created a data frame from each CSV, then started cleaning up all the five tables by finding if there were any duplicates present in the data if yes we dropped them and then we renamed all the columns of the data frame with meaningful names which would be helpful for us in writing queries later.

**Load:**

In the last part i.e. Load we established a connection with the MySQL server by creating an engine and then loaded all the five data frames into MySQL, before that we created a schema in MySQL in which all these five tables were loaded and we queried all these tables to check if our ETL project was working and we did get to run all the queries successfully.