**Database\_Task**

**Git** The data critical to our analysis was spread across four CSV files and all were sourced from [Kaggle](https://www.kaggle.com/).

The Chronicle of Higher Education provided us with data containing in-state and out-of-state tuition costs, fees, and other key features of universities such as whether the institution is public or private, the degree length offerings, and location.

Enrollment sizes along with school specific diversity statistics were obtained from [TuitionTracker.org](https://www.tuitiontracker.org/).

Potential salary data, percentage of STEM degrees and percentage of socially driven degrees comes from [payscale.com](https://www.payscale.com/).

The team was interested in determining if the US. Region or division might influence pay as well, so we utilized geographical groupings from the [United States Census Bureau](https://www.census.gov/geographies/reference-maps/2010/geo/2010-census-regions-and-divisions-of-the-united-states.html).

Next Slide

QuickDatabaseDiagrams was utilized to develop schema to create the database table structures and establish proper joins. Prior to importing the datasets, we went through a thorough cleaning exercise within Excel to standardize the University Name fields within each table. There were cases in which the same institution in each file had slightly different iterations; sometimes with articles, spaces, and/or symbols. It was critical for the University Name fields to align as it was the key in three of the four tables.

Once the extracted data files were cleaned and the design was developed, the datasets were loaded into a PostgreSQL database. SQL was utilized to join the university specific feature tables with the Census Bureau’s geographic table to provide a single output file used for our exploratory data analysis, machine learning models, and Tableau dashboard.