**Project Proposal – Group Microsoft**

**Team Members:**

* Jahangir Dewan
* Sofia Mbega
* Cody Miracle
* Lawson James
* Sajad Yarzada

**Project Description and Process Flow:**

The objective of the project is performing a complete ETL process using three different datasets and an API on United States unemployment rates, diversity index and median income:

1. **Extracting –** data from three different sources to show unemployment rates, diversity indexes and the median income for all US counties.
2. **Transformation –** finding the relevant variables, dropping unnecessary columns, and creating different relevant and useful tables
3. **Load –** Creating a database using MySQL and loading the Data Frames in a data base using SQLAlchemy or MongoDB.

**Datasets to Be Used:**

The following dataset will be used:

* US Counties Diversity Index - <https://www.kaggle.com/mikejohnsonjr/us-counties-diversity-index>
* US States and Counties Median Income - <https://data.world/tylerudite/2015-median-income-by-county>
* US States and Counties Unemployment - <https://www.kaggle.com/jayrav13/unemployment-by-county-us>
* US Bureau of Labor Statistics - <https://www.bls.gov/bls/api_features.htm>

**Rough Breakdown of Tasks:**

* Creating Github repository, branching and adding members
* Searching and finalizing datasets
* Start cleaning, analyzing, and visualizing the dataset using pandas as part of extraction and transformation process
* Loading the data frames in a SQL database
* Prepare final report/presentation