WORKFORCE PREPARATION ANALYSIS
ETL README

PROJECT 2

1. Project Background and Description

Evidence suggests that the national economy and several industries have been historically weakened or underserved by a lack of workforce talent. At the same time, public officials continue to question the value of higher education in relation to gainful employment. The present effort enhances our understanding of the labor market's social forces by accomplishing the following objectives executed through extracting, transforming, and loading (ETL) data:

1) Identifying major economic issues (national or international) in relation to industry;
2) Understand geopolitical sentiment of forces impacting the American labor market; and
3) Shed light on areas of higher education training that represent areas of weakness/misalignment and strength.

2. Scope of Methodological Work Within ETL Framework

Preliminary Activities. Prior to identifying relevant data sources, we analyzed documentation provided by the Economic Research Federal Reserve Bank of St. Louis. Specifically, we considered economic and marketplace research within FRED to accomplish Objective 1 (shown above). These areas were: immigration, trade, labor costs, and energy. Our initial interest in analyzing such broad and rich market forces was eventually tempered due to the unforeseen massive scope. The compromise was to select one major region, which was immigration. Consistent with said overzealousness and its essential compromise, fewer data sources were considered to ensure the project's manageability (see Project Proposal for a full listing of initial source considerations).

Data Sources, Extractions, Transformations, and Loadings. Primary information was extracted from the Bureau of Labor Statistics (BLS), the U.S. Census, the Integrated Postsecondary Education Data System (IPEDS), and the College Scorecard housed within the the U.S. Department of Education. Again, due to project limitations, labor market data were derived and only descriptively analyzed across labor force participation surveys. Descriptive statistics were inspected to come to know the spread of data. Said measures

revealed that variables required transformation, namely the age category and date and time objects as we explore recession influences on labor market participation rates. Data extracted from IPEDS were already prepared, as the Data System allows for a number of apriori specifications to maximize the extraction's usability and limit variable calls. Qualitative data were used from the U.S. Dept. of Education, representing another dataset extraction, to link program classifications to their respective career outcomes. Finally, five training programs with close ties to gainful employments impacting immigration – the initial major phenomenon – were pulled from major American colleges and universities were selected at random. The final piece of data, which was merged with IPEDS detail during the cleaning phase, was the College Navigator's rating information. Data were loaded into MongoDB.

3. Recommendations for Future Work

Future data scientists working in this area would benefit our understanding by considering the following:

1. Pull and merge a wider range of data available from the mentioned sources, as labor market phenomena and market force connections are difficult to unpack without richer data.
2. Analyze variables both descriptively and inferentially to unpack associations relevant for this work. At the same time, consider factor analyzing variables to create stronger constructs and, therefore, generalizations.
3. Scrape social media data to verify the truthfulness of the career outcome detail identified in the U.S. Dept. of Education's crosswalk exposition.