**Ryan McConnell – Data Profiling Project**

**Approach and Progress –**

* Quick overview of 5 individual datasets, understanding referential links.
* Create ‘Profile Summary’ tabs in each .csv to get a high-level view of aggregate data records, records with compete data by field, and also records without data by field, as well as a record sample of each field.
* Look for common format issues by quick import of data files into Access to create tables. I did this because Access’s automatic Import Error summary tables can quickly identify problems like a\_\_company’s ‘zipcode’ field not in the correct format for importing into an Access table.

**Next Steps –**

* Continue reviewing data formats in remaining datasets and data completion. Explore opportunities to update the business process of storing data in correct formatting as data is extracted from source.
* Compile a list of field exclusions to present to business owners (i.e. Do fields like ‘fax’ in a\_\_company need to be retired? Are new company records being created with fax numbers?) to reduce integration workloads.
* Identify and investigate duplicate address/address\_id, country/iso\_country, and zipcode/location\_postal\_code and explore possibility for mapping relations between the a\_\_ and b\_\_ datasets where duplicates do not occur.