1. **DATA COLLECTION & PREPARATION**
   1. **Model Application – rank\_ppd\_polo\_addrs.py**
      1. Collect AIMS tables:
         1. entity\_comm\_usg\_at
         2. post\_addr\_at
         3. license\_lt
         4. entity\_key\_et
      2. Collect current PPD, ppd\_df
      3. Load saved model and model variables
      4. DATA PROCESSING – pull database and ppd info, format for model
         1. **create\_ppd\_scoring\_data**(ppd\_df, ppd\_date, ent\_comm\_df, ent\_comm\_usg\_df, post\_addr\_df, license\_df, ent\_key\_df)
            * remove records with null line 2, state, or zip
            * filter to DPC (‘TOP\_CD’ == ‘020’)
            * date\_df = ppd\_df[[‘ME’, ‘ppd\_date’]
            * **entity\_df = create\_addr\_entity\_data**(ent\_comm\_df, ent\_comm\_usg\_df, post\_addr\_df, license\_df, ent\_key\_df, date\_df)

**clean\_ent\_comm\_data(**ent\_comm\_df**)**

**clean\_post\_addr\_data(**post\_addr\_df**)**

**clean\_ent\_usg\_data(**ent\_comm\_usg\_df**)**

(all these functions do is reduce the df to a subset of columns)

* + - * + ppd\_df = **create\_addr\_key(**ent\_comm\_df, ent\_comm\_usg\_df, post\_addr\_df, license\_df, ent\_key\_df**)**
        + ent\_ppd\_df = ppd\_df.merge(entity\_df, on=’ME’)
        + ppd\_final\_df = **rename\_ppd\_columns**(ent\_ppd\_df)

literally just 1 function that uses a big dictionary to insert a “ppd\_” prefix before all PPD column names.

* 1. **Model Training – Preprocess\_PoloRank\_ModelData.ipynb**