ReadMe

Data Details

* “Pfizer Turnover Prediction Model.csv” (4,723,161 rows x 59 columns)
* *Status* = 1 indicates where colleague has voluntarily left the company. *Status* = 0 is for all other colleagues. There are 4628 rows with *Status* = 1
* Columns *POS\_LEVEL1, POS\_LEVEL2, POS\_LEVEL3, POS\_LEVEL4, and POS\_LEVEL5* are character columns with leading zeroes. Explicitly specify them as such to avoid dropping the leading zeroes
* Fields with prefix “County Level” are aggregated at a country level

Geospatial Matching

* Currently, fields with prefix “Country Level” are aggregated at a country level. To get a more granular level of detail these same fields can be aggregated a city/region level. These can be brought in from the following files:
  + “demand\_pressure.xlsx”
  + “location\_parameters.xlsx”
* In the “master\_output\_anonymized.csv” main file there is a *Location* field. You can match this against the *City* field in the aforementioned two files. To do this I would recommend the following procedure:
  + Obtain a list of cities from an open source dataset
  + Use levenshtein distance or comparable method to match *Location* against the open source city list and clean the *Location* field such that it only coutnains the city name
  + Not all cities in the *Location* field will obviously match up against the *City* field. To get the best possible match, use a geospatial proximity match to match *Location* to the nearest *City* by x, y coordinate distance