HW1

1.Write code to evaluate the results of a missing value imputation algorithm on the missing data file (from HW0), in terms of the average absolute error (i.e., difference between the actual value and the imputed value) over all the ten data points (faculty members) in your HW0.

2.Try the following alternatives, possibly using the impute.jl Julia package (https://github.com/invenia/Impute.jl), any other publicly available code, or your own code:

(a) Individual_mean: Replace by the mean over all years (for which data is available) for that faculty member.

(b) Individual_median: Replace by the median over all years (for which data is available) for that faculty member.

(c) Field_mean: Replace by the mean over all faculty members (for whom data is available) for that year.

(d) Field_median: Replace by the median over all faculty members (for whom data is available) for that year.

(e) Local_gradient: For missing data in the middle years (2018-2021), replace by the average of the preceding and following years, for that faculty member.

For missing data in 2017, use twice the value of 2018 minus the value of 2019. For missing data in 2022, use twice the value of 2021 minus the value of 2020.

(f) Nearest_neighbor_L1: Calculate an averaged distance from each of the other faculty members, based on fields for which both faculty members have values, e.g., the L1 distance between (..., n17, , n19, ,n20, n21, n22, ...) and (..., m17, m18, ,m20, m21, m22, ...) = (|n17-m17| + |n20-m20| + |n21-m21| + |n22-m22|)/4. For each faculty member, compute the nearest other faculty member using this distance measure, for whom the missing value is not in the same field, and replace the missing value by that field value for the nearest neighbor.

(g) Nearest_neighbor_L2: Similar to the above but use the Euclidean distance measure, e.g., the L2 distance between (..., n17, , n19, ,n20, n21, n22, ...) and (..., m17, m18, ,m20, m21, m22, ...) = ((n17-m17)^2 + (n20-m20)^2 + (n21-m21)^2 + (n22-m22)^)2/4.

3. Compare the results of the different approaches. Which approach has worked best?