I started with the data sets I created after cleaning and shaping the data (HW1), that is why the ‘Output.csv’ is the input.

First, I found all the states that have missing values through python, which are 'California', 'Colorado', 'Georgia', 'Hawaii', 'Indiana', 'Louisiana', 'Minnesota', 'Oklahoma'. Then I decided to use different strategies, LCOF (ffill / pad), backfill and mean, to add appropriate values.

Here are the full explanations for the strategies used for each state:

Since we only have the divorce rate of 1990 in California, "backfill" method is the optimal one because the last-known data of divorce rate is available at the following time point. Similarly, the "backfill" strategy can also be use in refilling the missing data of Georgia, Hawaii, and Minnesota since they all share the same missing pattern.

For Colorado, since only one value was missing, it is sufficient to refill the value by the mean of the preceding and following years. There is no single value for the Divorce rate of Indiana, so the best way to refill the blank is to use the overall mean.

It is sufficient to use the forward refilling strategy to refill the NA of Oklahoma. For Louisiana, both the forward and the backward refilling strategies are used.