For the crime dataset we chose to split the available dataset into 80% training and 20% testing. The rationale for this decision was that our dataset is fairly large, and we don’t need extra training data points to increase the accuracy of the prediction. Moreover, the 80/20 split ratio is a common practice in the data science literature which is another reason why we decided to follow this community guideline. Another reason why we think the model has a good predicting power is that the data fits into the model quite well which was calculated by using Cross-Validated R2 and Cross-Validation RMSE for each of the 10 recommended by the literature folds:

Cross-Validated R2 For All Folds: 0.767

Cross-Validation RMSE For All Folds: 18686.9279

After The visualization it is easy to notice that for states that were missing the values now the data looks like it doesn’t follow the pattern. This is especially visible in the state of New York where in the period 1960-1965 there is a big spike in the crime rate that looks like it doesn’t belong there. Here, it is important to mention that this is likely due to nature of the linear regression models where the model is a straight line through the whole dataset. Due to this nature some of the features of the dataset like local rises and falls in the value might be lost which might result in data being “locally unfit” but fit well within a dataset as a whole.

Our Tableau visualization was chosen to show the differences in the data sets before and after replacing the missing values. The first set of graphs show the general difference of the data over the years, while the animated set of graphs show the differences for each state, which allows the user to view more closely the altered data and the data values that were placed by the python program.

https://public.tableau.com/profile/melissa.oshiro?fbclid=IwAR30yo6SbAA0St7g2zHdIUf3hLvM6VeqXf72AOKZjwTxfukB6OdVp6Mipws#!/vizhome/hw5-handlingmissingdata/HandleMissingData?publish=yes