Name: Heman Baral
CMSC 435

Assignment 2

1. **What are the MAE values for the eight results?**

```
MAE_05_mean = 0.0639
MAE_05_mean_conditional = 0.0629
MAE_05_hd = 0.0361
MAE_05_hd_conditional = 0.0363

MAE_20_mean = 0.0645
MAE_20_mean_conditionalal = 0.0633
MAE_20_hd = 0.0435
MAE_20_hd_conditional = 0.0437
```

2. **Which of the considered four imputation methods is the least accurate for the dataset_missing05.csv dataset? Briefly explain why this method is worse than the other three methods.**

   For dataset_missing05 the least accurate imputation method was Mean Imputation. In this method we are calculating the mean of entire column and replacing each missing value with the computed mean. This method reduces the variance of the imputed variables. Mean imputation treats each missing value carries the same weight which result to unreal and bias value.

3. **Are the results for the same algorithm on the two data sets the same/worse/better (e.g. mean on dataset_missing05 vs. mean on dataset_missing20)? Briefly explain why.**

   mean:

   Dataset_missing20 value is about the same than Dataset_missing05. Mean for the dataset_missing05 and dataset_missing20 were about the same which result in similar mean absolute error.

   mean conditional:

Dataset_missing20 value is about the same than Dataset_missing05. Mean conditional for both class (yes and no) for dataset_missing05 and dataset_missing20 were similar.

hot deck:

Dataset_missing20 value is worse than Dataset_missing05. Hot deck imputation is better but there is a major difference with 5% and 20% missing value. Hot deck makes implicit assumptions through the choice of metric to match near neighbor value and missing value. The more data is missing there will more assumption which can result for MAE to increase.

hot deck conditional:

Dataset_missing20 value is worse than Dataset_missing05. The hot deck conditional works the same way as hot deck, the difference is first we compute the distance between two values with similar features. As the missing value increases the amount of error to be occur also increases.

4. **Which of the two unconditional methods (mean vs. hot deck) is faster, i.e., requires fewer computations? Briefly explain why. Give computational complexity of both methods as a function of the number of objects n and use it to support your explanation.**

   Mean imputation was the fastest because it has to read all the data in a column once to compute the mean and replace the missing value whereas hot deck has to look for the nearest distance neighbor on every single missing value is encountered.

   Complexity:

   Mean Imputation: O(n^2)

   Hot Deck Imputation: O (n^3)