Q7)

|  |  |  |
| --- | --- | --- |
|  | Validation Set MAE | Test Set MAE |
| Drop columns with missing values | 17837.82570776256 | 16381.48041 |
| SimpleImputer, strategy=median | 17791.59899543379 | 16452.43726 |
| Add new colums, SimpleImputer, strategy=median | 18063.910194063923 | 16451.15081 |

Q8)Which approach gives the best validation set score? Which approach gives the best test set score (i.e. leaderboard score)? If the approach that gives the best validation score is different from the approach that gives the best test set score, what could be the reason for this behavior?

* SimpleImputer, strategy=median gives the best validation set score.
* Drop columns with missing values gives the best test set score.

If the approach has a better score in the validation set it may be due to overfitting. This event occurs when we have a complex model, or the model learned the training data well and is incapable/less capable of new data. Additionally, this can happen validation set is not representative enough.