Output file created by the reducer was renamed to add an extension, .txt, to allow it to be opened in Jupyter notebook.

Decision tree and KNN based imputation methods are appropriate for non-parametric data such as this [1].

The data is assumed to be MCAR, as there is no known reason why there would be no tweets at all on the missing dates, and there is no common factor amongst the missing data [2].

[1] Bertsimas, D., Pawlowski, C., and Zhuo, Y. D. (2018). From Predictive Methods to Missing Data Imputation: An Optimization Approach. *Journal of Machine Learning Research*, 18, pp.1-39.

Yang, L. and Chiang, J.A. (2020) ‘Use case and performance analyses for missing data imputation methods in big data analytics’, *Proceedings of 2020 6th International Conference on Computing and Data Engineering* [Preprint]. doi:10.1145/3379247.3379270.

To use ycsb with Cassandra a change was needed to conf/cassandra.yaml, after [<https://github.com/brianfrankcooper/YCSB/pull/98/files>]

To run the Cassandra load test in ycsb a ycsb keyspace had to be created, as shown in appendix, and similarly a usertable had to be created in hbase.

‌