Missing Data: Introduction

Exploratory data analysis is mostly about gaining insight through visualization and hypothesis testing. Recall from the beginning of this unit that the ETL process is useful for holding the data to a minimum standard with respect to quality assurance. This unit deals with the imputation of missing values and it is where EDA and ETL meet. Missing value imputation could exist as part of the ETL process, but it is not often clear which strategy is the best until we can make comparisons. The comparisons are best made by evaluating model performance using a hold-out data set. One missing value strategy may be better for some models, but for others another strategy may show better predictive performance.

Our Story

Missing data is a common problem in most real-world scientific datasets. While the best way for dealing with missing data will always be preventing the occurrence in the first place, it will still remain a problem. Sometimes data is collected from sensors that fail to record or data collection is distributed across individuals and the merged data does not harmonize well. There are a variety of ways for dealing with missing data, from more simplistic to very sophisticated, but a standard metric by which we measure utility will still be model performance.

This module is focused on ensuring that you are aware of how to deal with missing values and the consequences that arise from deciding how they are dealt with. At AAVAIL and nearly all companies with accumulated data there eventually you will encounter missing values and the actions you take at this stage of the overall workflow can heavily influence the downstream business of your models utility.

THE DESIGN THINKING PROCESS

This unit relates to the *Empathize* phase of their design thinking process. Discuss with your colleagues to ensure that you understand why data are missing. It is not enough to systematically deal with missing values in the same way for every dataset you encounter. Significant model performance improvements can be achieved by leveraging discussions with with those close to the data generation process.