# Evaluation of Imputation Methods Under Different Missing Data Conditions

Yehchan Yoo, Department of Statistics

## INTRODUCTION

### Missing data is common in real-world survey data and is becoming more of a problem everyday.

- More than 1/3 of the data in prevalent surveys in U.S., U.K., Mexico, Taiwan, and Japan found to be missing according to Dodeen (2018)
- U.S. federal surveys suffering from decreasing survey response rates as of 2016 (Czajka & Beyler, 2017)
- From here, various **imputation** methods can be used to fill in the missing data and reduce bias (Lohr, 1999, pp. 277-278).

### Types of Missing Data Conditions (Mack et al., 2018)

- **MCAR (Missing Completely At Random):** Missingness independent of observed and unobserved data
- **MAR (Missing At Random):** Missingness independent of unobserved data but not independent of observed data
- **MNAR (Missing Not At Random):** Missing data not independent of either unobserved or observed data

### Goal: Simulate missing data in a given dataset and evaluate the performance of different imputation methods.

## DATA

### American Community Survey

- 1-year ACS Public Use Microdata Sample (PUMS) dataset from New York state from 2023 (United States Census Bureau, 2024)

AMERICAN COMMUNITY SURVEY
U.S. CENSUS BUREAU

NEW YORK STATE

## METHODOLOGY

### Steps

**Target variables:** `VALP` (property value) and `RNTP` (contract rent)

1. Get a subset of the ACS NYS data with non-missing `VALP` values and another subset of the data with non-missing `RNTP` values.
2. Simulate MCAR, MAR, and MNAR conditions for *only* the corresponding target variable (i.e. `VALP` for the `VALP` subset, `RNTP` for the `RNTP` subset) in each of these subsets.
3. Impute the missing data *only* in the corresponding target variable using the imputation methods below.
4. Calculate the means and quartiles of the target variables with corresponding standard errors using the replication weights provided in the dataset.

### Simulating Missing Data

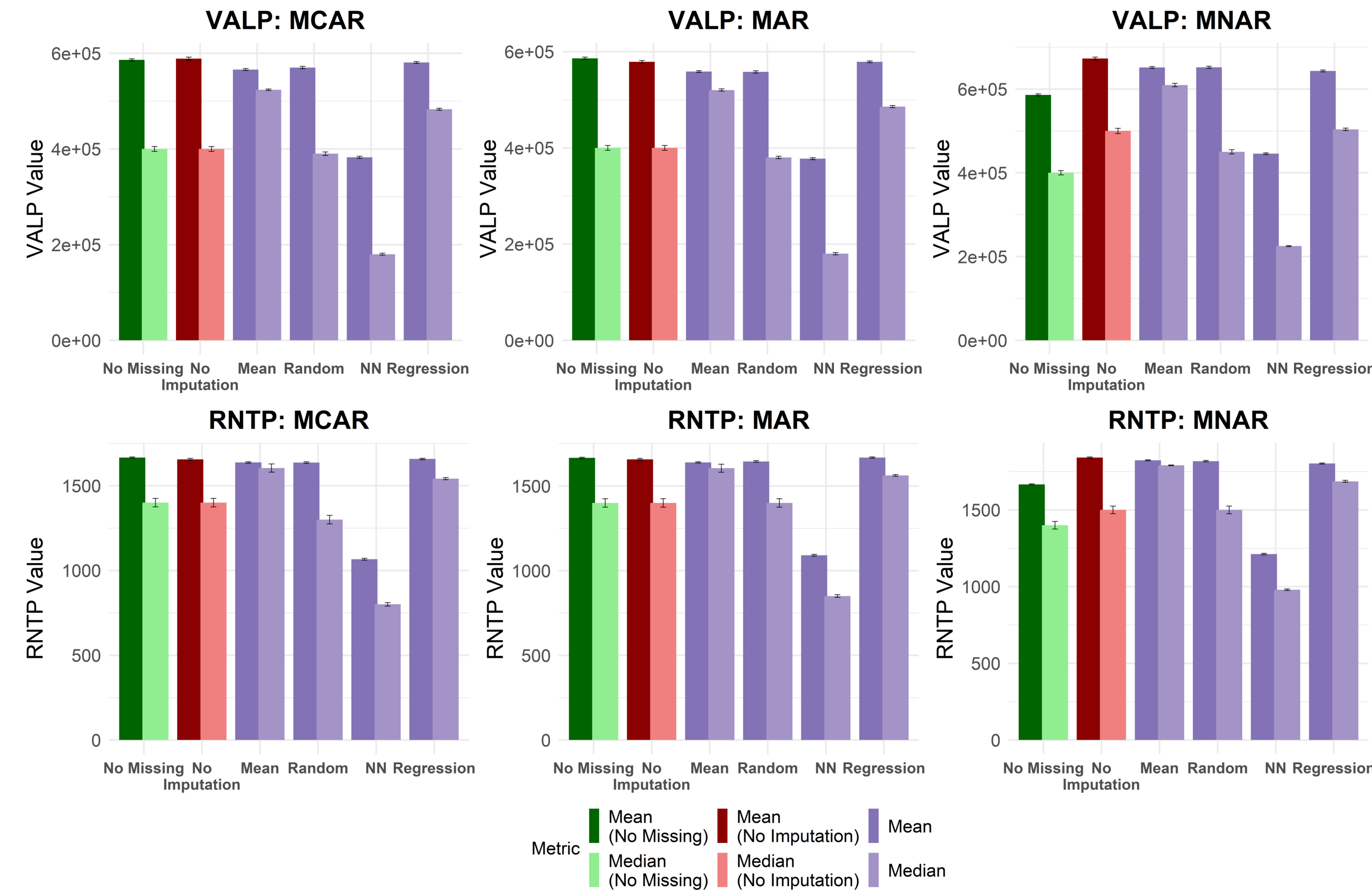
Aimed for 35% of the data in the corresponding target variable to be missing & followed the simulation guidelines of Zhang (2021):

- **MCAR:** Each value missing with 35% chance
- **MAR:** Missingness based on mode of response (online vs. offline) to reflect real-world pattern in response rate by response mode (Shiyab et al., 2023)
- **MNAR:** Missingness based on the values of the target variable itself to reflect higher nonresponse rate for lower income/house value in the real world (Peterson et al., 2021)

### Imputation Methods (Lohr, 1999, pp. 274-276)

| Control | Hot-Deck | Classical |
|---|---|---|
| • No missing (before simulating missingness)<br>• No imputation (after simulating missingness) | • Random Imputation (Value from randomly picked row with non-missing value)<br>• Nearest Neighbor (NN) Imputation (Mean value from three closest matching cases) | • Mean Imputation (Fill with mean value)<br>• Regression Imputation (Predict values to fill in missing values based on other columns using regression model) |

## RESULTS & DISCUSSION




- Bias from nearest-neighbor imputation consistently large across all missingness types for both target variables, vastly underestimating both means and medians
- Bias from random imputation consistently low across all missingness types for both target values, especially for median estimation
- Having no imputation leads to estimates close to the baseline ("No Missing") under MCAR and MAR but noticeably biased under MNAR due to loss of low-value entries
- Standard errors generally relatively high under mean and random imputations for median estimation compared to other imputation methods
- Mean and regression imputations distort median estimates by oversimplifying the distribution of missing values and by not accounting for the skewedness of the target variables

## NOTE

This project was done for STAT 529: Sample Survey Techniques under the guidance of Professor Robin Mejia.

**References (Poster Only):** **Czajka, J. L., & Beyler, A. (2017).** *Declining response rates in federal surveys: Trends and implications* (tech. rep.) (Accessed: 2025-05-21). Office of the Assistant Secretary for Planning, Evaluation, U.S. Department of Health, and Human Services. https://aspe.hhs.gov/sites/default/files/private/pdf/255531/Decliningresponserates.pdf; **Dodeen, H. (2018).** The prevalence of missing data in survey research. *International Journal for Innovation Education and Research*, 6. https://doi.org/10.31686/ijier.vol6.iss3.978; **Lohr, S. L. (1999).** *Sampling: Design and analysis*. Duxbury Press.; **Mack, C., Su, Z., & Westreich, D. (2018, February).** Types of missing data. https://www.ncbi.nlm.nih.gov/books/NBK493614/; **Peterson, S., Toribio, N., Farber, J., & Hornick, D. (2021, March).** *Nonresponse bias report for the 2020 household pulse survey* (tech. rep). United States Census Bureau. https://www2.census.gov/programs-surveys/demo/technical-documentation/hhp/2020_HPS_NR_BiasReport-final.pdf; **Shiyab, W., Ferguson, C., Rolls, K., & Halcomb, E. (2023).** Solutions to address low response rates in online surveys. *European Journal of Cardiovascular Nursing*, 22 (4), 441–444. https://doi.org/10.1093/eurjcn/zvad030; **State of New York. (2015).** New york state wordmark [Public domain image; may be subject to trademark restrictions]. https://commons .wikimedia.org/wiki/File:New_York_State_wordmark.svg; **United States Census Bureau. (2008).** American community survey logo [Public domain image from the U.S. federal government]. https://commons.wikimedia.org/wiki/File:US-Census-ACSLogo.svg; **United States Census Bureau. (2024).** American Community Survey (ACS) 2023 1-Year Public Use Microdata Sample (PUMS) [Accessed: 2025-05-21]; **Zhang, X. (2021, July).** Tutorial: How to generate missing data for simulation studies [Retrieved from https://doi.org/10.20982/tqmp.19.2.p100]. https://doi.org/10.20982/tqmp.19.2.p100