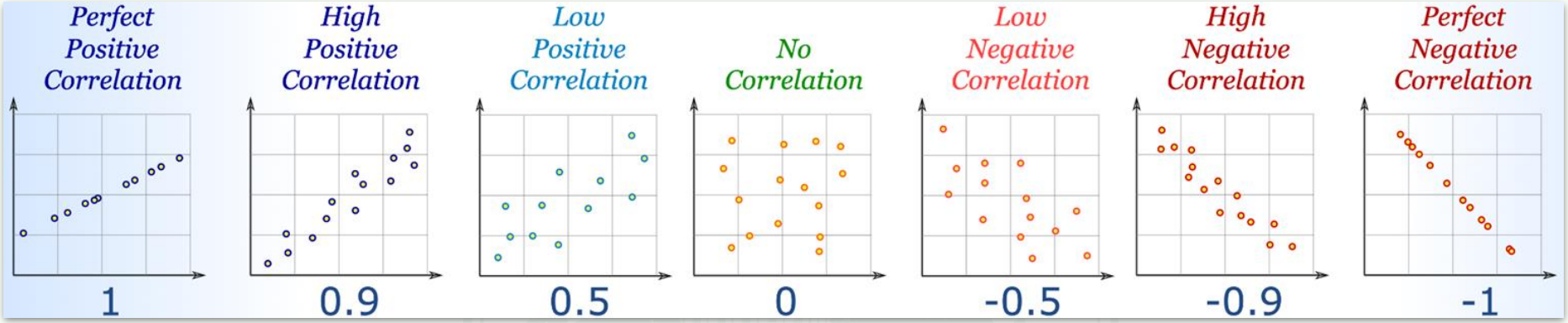
Missingness

Prof. Murillo
Computational Mathematics, Science and Engineering
Michigan State University

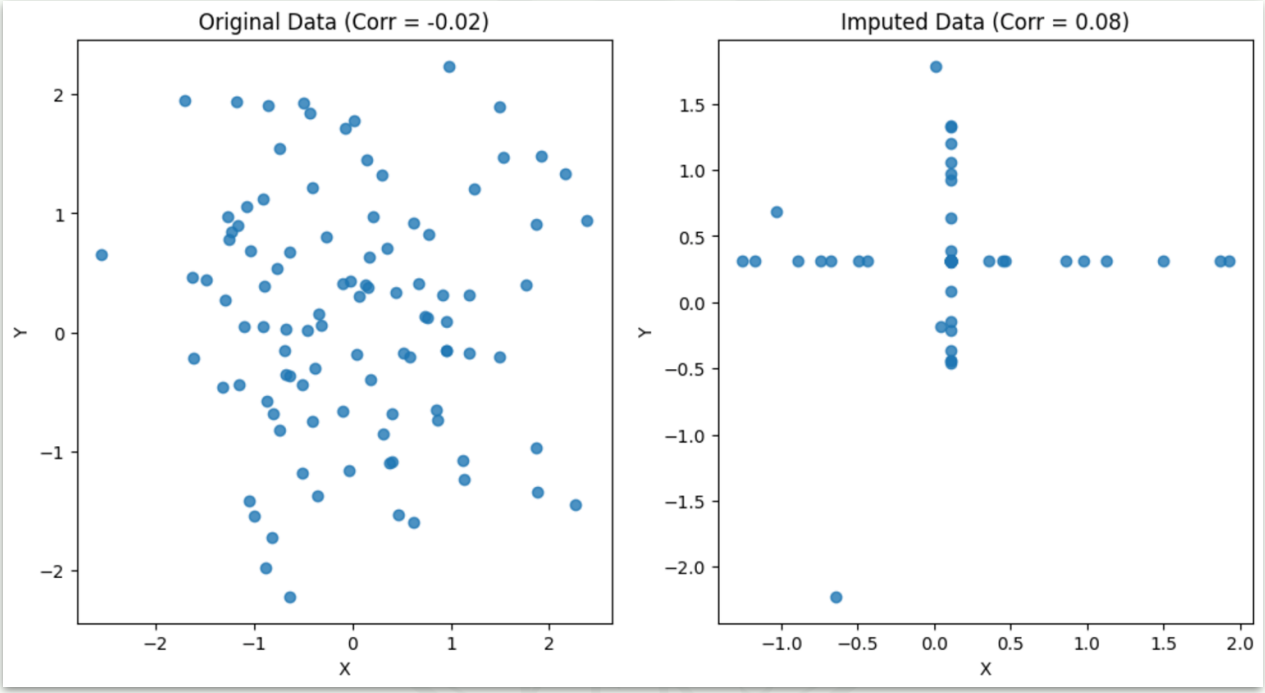
Loss of Correlations from Mean Imputation



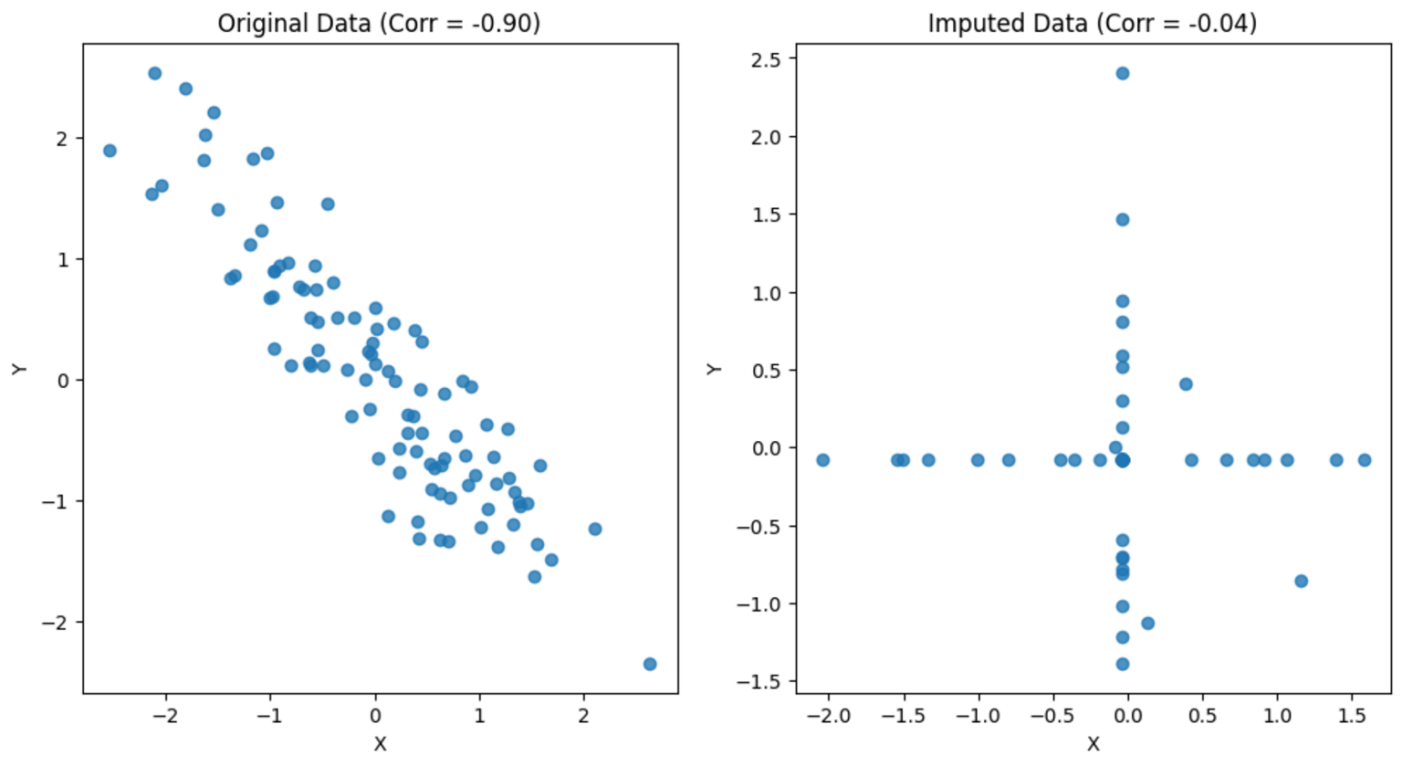
$$\rho_{X,Y} = \frac{\mathbb{E}[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$

Loss of Correlations from Mean Imputation: Examples

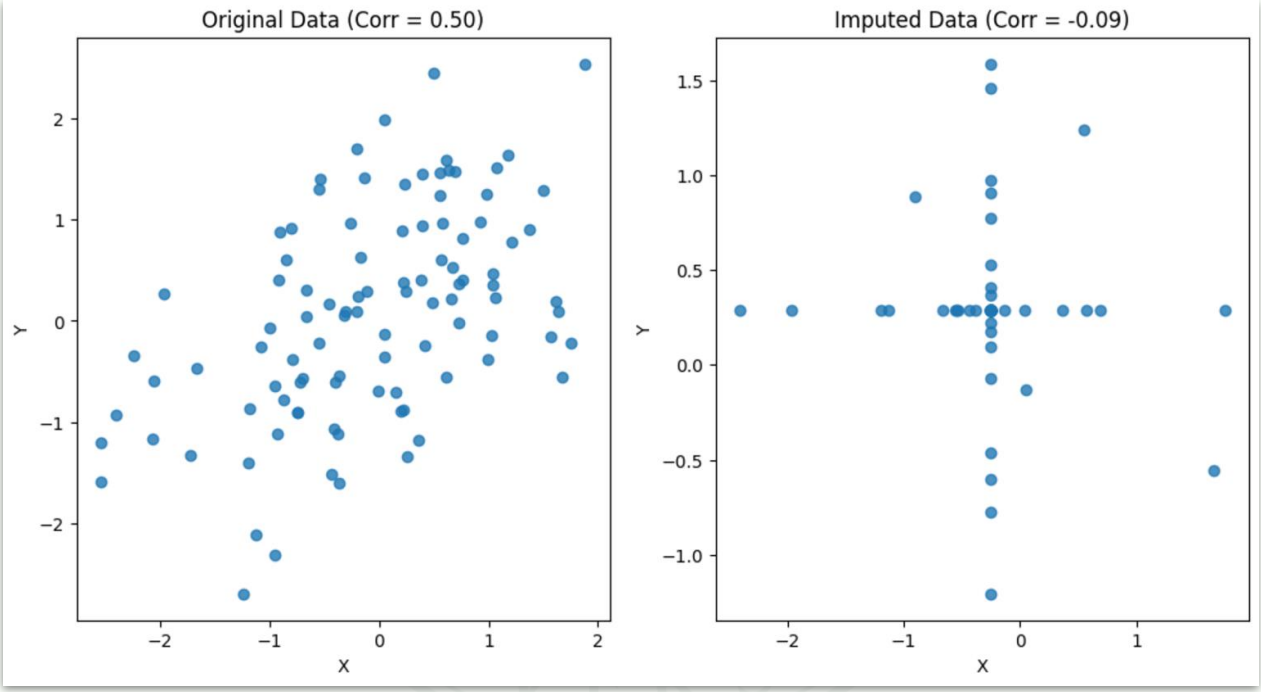
missing rate = 0.8



Loss of Correlations from Mean Imputation: Examples



Loss of Correlations from Mean Imputation: Examples



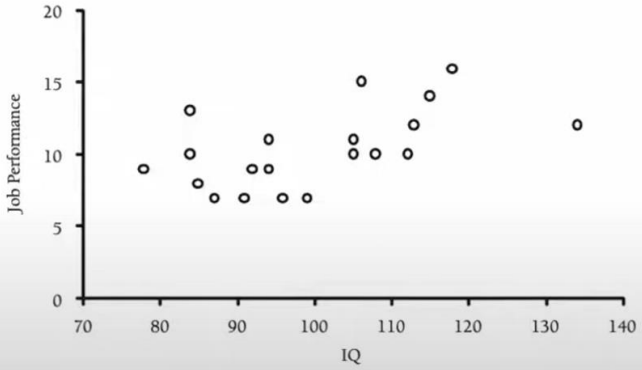
But First....Three Topics From Last Time

- impact of mean imputation on correlations
- stochastic regression
- diagnosing missingness
- visualization

Stochastic Regression: What is the Goal?

Here is the raw data:

Example dataset



2.1. Complete-data scatterplot of the IQ and job performance scores from Table 2.1.

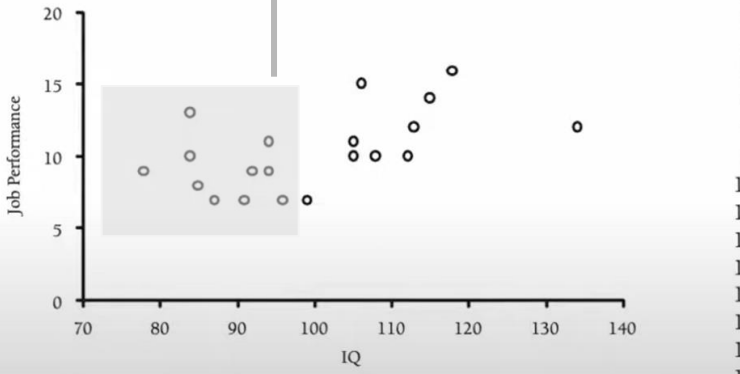
TABLE 2.1. Employee Selection Data Set

Complete data		Missing data
IQ	Job performance	Job Performance
78	9	—
84	13	—
84	10	—
85	8	—
87	7	—
91	7	—
92	9	—
94	9	—
94	11	—
96	7	—
99	7	7
105	10	10
105	11	11
106	15	15
108	10	10
112	10	10
113	12	12
115	14	14
118	16	16
134	12	12

Stochastic Regression: What is the Goal?

Here is the synthetic data:

Example dataset

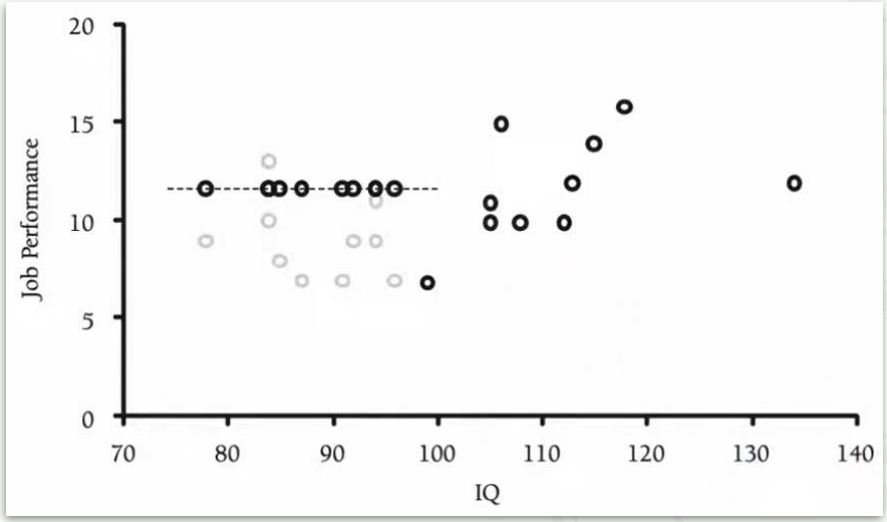


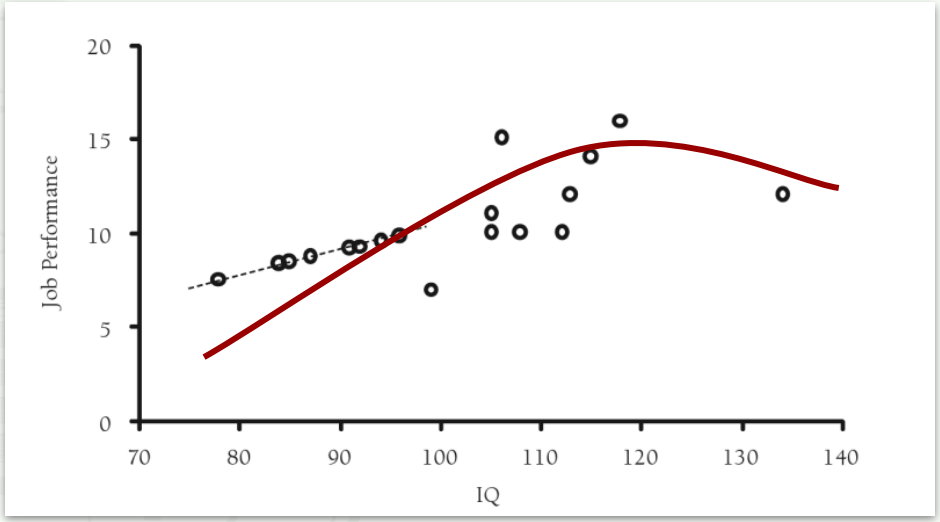
2.1. Complete-data scatterplot of the IQ and job performance scores from Table 2.1.

TABLE 2.1. Employee Selection Data Set

Complete data		Missing data
IQ	Job performance	Job Performance
78	9	—
84	13	—
84	10	—
85	8	—
87	7	—
91	7	—
92	9	—
94	9	—
94	11	—
96	7	—
99	7	7
105	10	10
105	11	11
106	15	15
108	10	10
112	10	10
113	12	12
115	14	14
118	16	16
134	12	12

Fit Data and Extrapolate



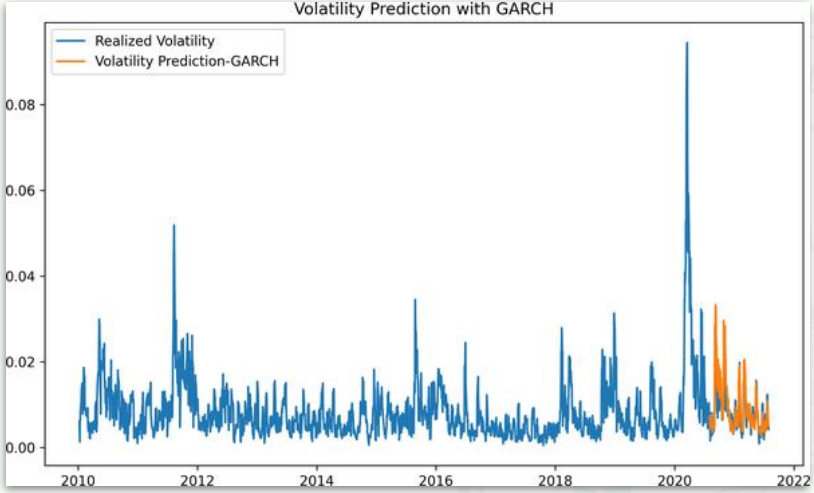


Point #1: mean imputation should “never” be done

Point #2: fitting is better, and perhaps good enough?

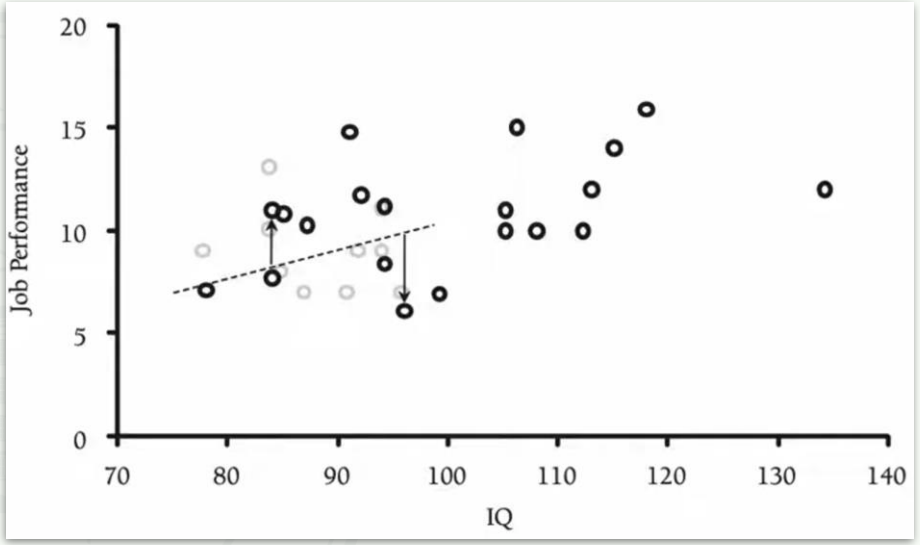
Point #3: don’t need to use a line (first-order polynomial)

Variance/Volatility



Very often we want to extrapolate the "volatility" in our data.

Sometimes, we only have "volatility".



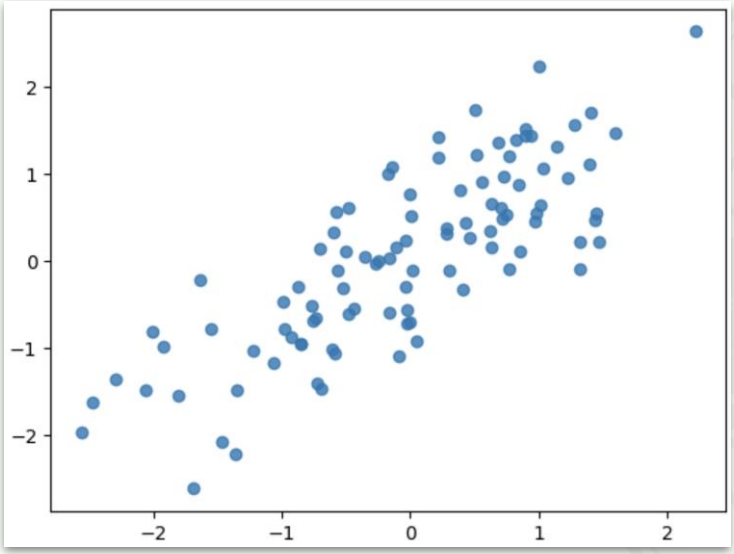
Point #4: when we impute, we want to preserve the mean, trend and variance

multiple imputation can change the conclusion!

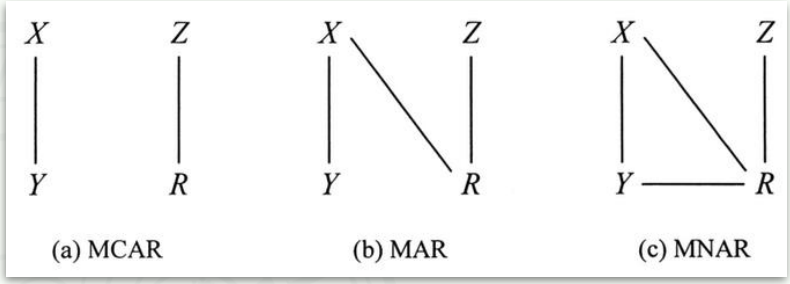
But First....Three Topics From Last Time

- impact of mean imputation on correlations
- stochastic regression
- diagnosing missingness
- visualization

More Details on Missingness



original




```
19 # Introduce MCAR missingness
20 missing_rate = 0.2  # Proportion of missing values
21 df_mcar = df_original.copy()
22 for col in df_mcar.columns:
23     missing_indices = random.sample(range(n), int(missing_rate * n))
24     df_mcar.loc[missing_indices, col] = np.nan
25
26 # Introduce MAR missingness
27 df_mar = df_original.copy()
28 missing_indices_mar = df_original[df_original['X'] > df_original['X'].quantile(0.75)].index
29 df_mar.loc[missing_indices_mar, 'Y'] = np.nan
30
31 # Introduce MNAR missingness
32 df_mnar = df_original.copy()
33 missing_indices_mnar = df_original[df_original['Y'] > 0].index
34 df_mnar.loc[missing_indices_mnar, 'Y'] = np.nan
```

add missingness to the data.

Missingness in Terms of Conditional Probabilities

MCAR:

$$P(\text{missing}|\text{complete}) = P(\text{missing})$$

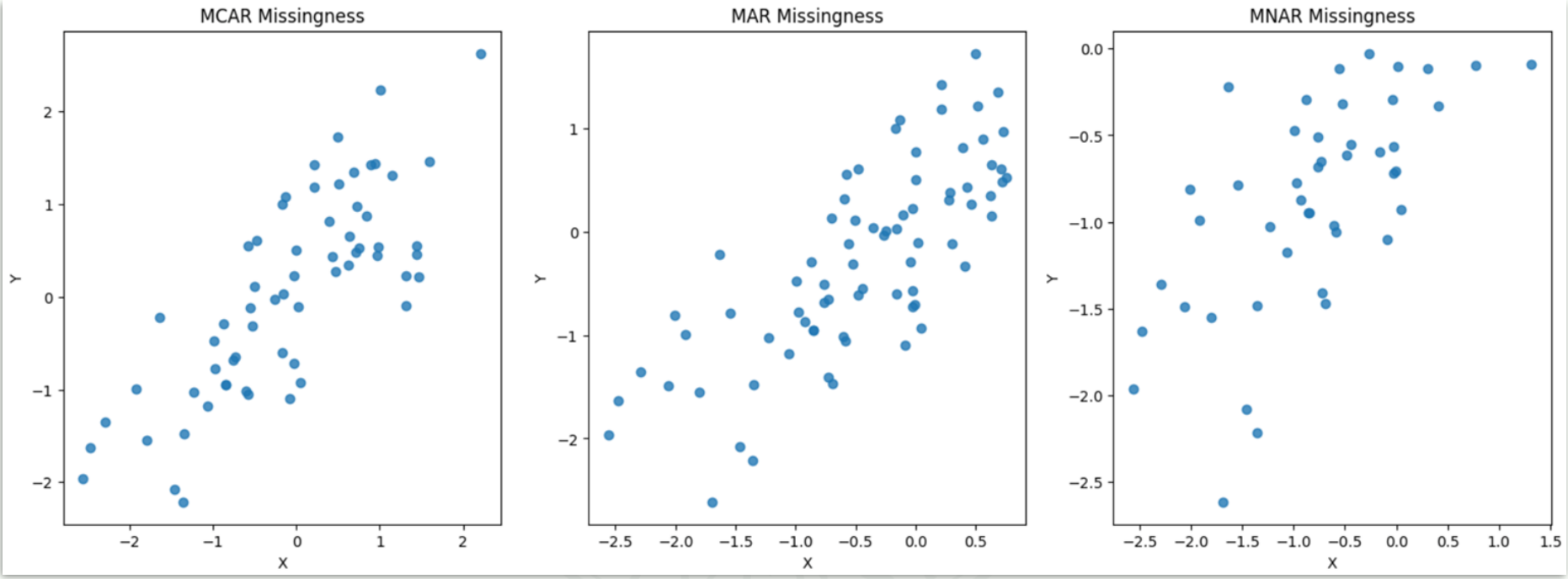
MAR:

$$P(\text{missing}|\text{complete}) = P(\text{missing}|\text{observed})$$

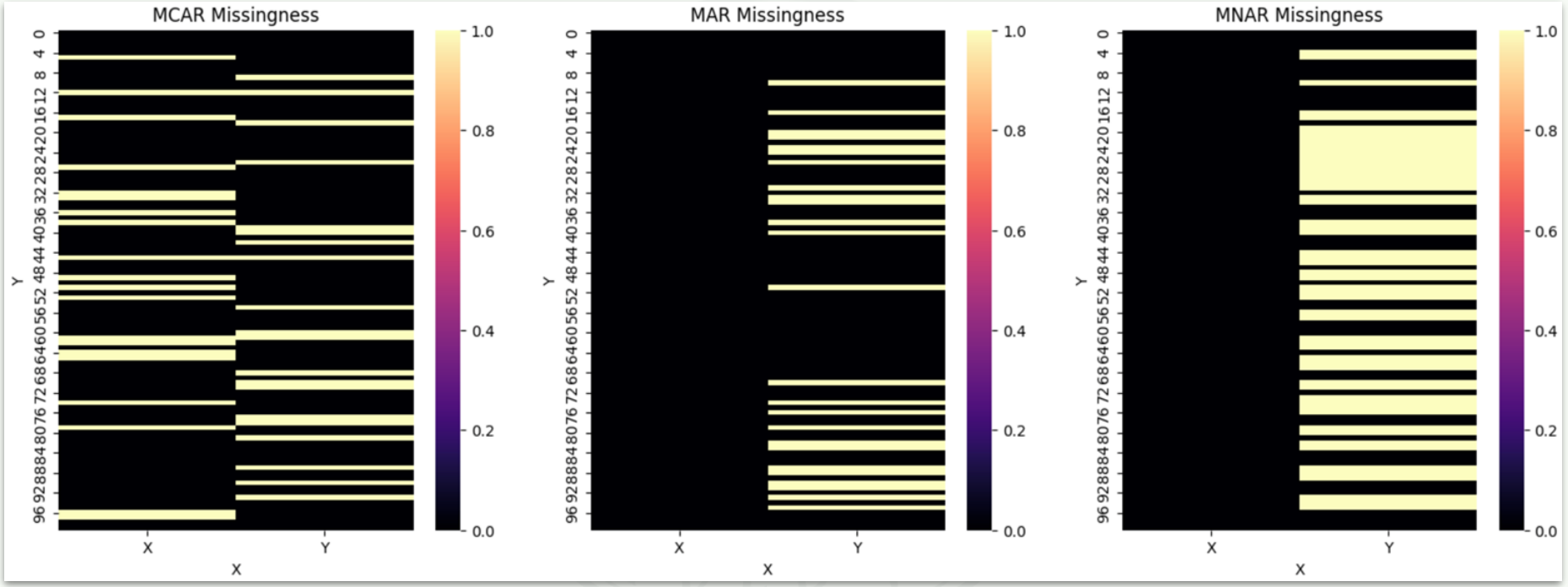
MNAR:

$$P(\text{missing}|\text{complete}) \neq P(\text{missing}|\text{observed})$$

More Details on Missingness



Visualizing Synthetic Missingness



Sorted by X Value