HANDLING MISSING VALUES

Prof. Dr. Dyah Erny Herwindiati. S.Si. M.Si.
Lely Hiryanto. S.T.. M.Sc.. Ph.D.

Missing Values

- Banyak dataset real yang data observasinya tidak lengkap (tidak diketahui nilainya atau ‘hilang’):
  - di satu variable beberapa observasi tidak ada atau bisa tidak ada sama sekali.
  - Untuk data seperti nama, alamat atau email dari sejumlah pelanggan kosong.
- Salah satu pendekatan yang paling sederhana adalah tidak menggunakan setiap data observasi yang kosong tersebut
- Jika training data set berjumlah sedikit dan kita membutuhkan semua data tersebut maka perlu melakukan estimasi untuk data yang hilang
  - Untuk **data bertipe kategorikal**, estimasi data yang hilang bisa menggunakan **data mayoritas** dari data pelatihan yang memiliki nilai target yang sama dengan data yang hilang tersebut
  - Untuk data numerical, dapat di-estimasi menggunakan rata-rata data pelatihan yang memiliki nilai target yang sama dengan data yang hilang tersebut

Data Missing Mechanism

- Missing Completely At Random (MCAR)
  - Nilai/data yang hilang tidak memiliki hubungan atau pola dengan nilai lainnya dalam satu variabel yang sama atau dengan variabel lainnya
- Missing At Random (MAR)
  - Nilai yang hilang bisa diprediksi berdasarkan variable lainnya, tapi bukan dari nilai yang hilang tersebut
- Missing Not At Random (MNAR)
  - Sejumlah nilai untuk sebuah variable hilang karena kondisi dari variabel tersebut dan tidak tergantung dari nilai-nilai variabel lainnya

Complete Dataset		MCAR	MAR	MNAR
IQ	Ratings	Ratings	Ratings	Ratings
78	9	?	?	9
84	13	13	?	13
84	10	?	?	10
85	8	8	?	?
87	7	7	?	?
91	7	7	?	?
92	9	9	9	9
94	9	9	9	9
94	11	11	11	11
96	7	?	7	?
99	7	7	7	?
105	10	10	10	10
105	11	?	11	11
106	15	15	15	15
108	10	10	10	10
112	8	?	8	?
113	12	12	12	12
115	14	14	14	14
118	16	16	16	16
134	12	?	12	12

Methods for Handling Missing Values

- **Deletion**
  - Pendekatan yang sederahan dan mudah untuk dilakukan
  - Sesuai untuk MCAR
  - Metode: Listwise and Pairwise
  - Kerugian: kehilangan data yang bernilai dan dapat menghasilkan hasil analisis yang bias.
- **Imputation**
  - Mendapatkan nilai untuk data yang hilang dan secara bersamaan mempertahankan akurasi dan konsistensi keseluruhan nilai di dataset
  - Metode: statistical based imputation, machine learning, and neural network
- **Interpolation**
  - Estimate berdasarkan nilai data yang ada
  - Cocok untuk data time series (nilai yang hilang diasumsikan mengikuti sebuah pola tertentu dari data pelatihan
  - Metode: piece wise interpolation, linear interpolation, polynomial interpolation, and spline interpolation
- Representation Learning
  - Ekstraksi fitur dari dataset asli (raw dataset) dengan menemukan struktur dan pola dari dataset tersebut
  - Metode: Graph Neural Network and AutoEncorders

Deletion

- **Listwise**: menghapus semua observasi yang memiliki satu atau lebih nilai yang hilang
- Pairwise: Sebuah variable dengan data yang tidak lengkap tidak disertakan dalam analisis yang membahas hasil observasi dari variable tersebut
  - Contoh: jika analisis membahas data ratings, maka analisis tersebut tidak dilakukan

IQ	Ratings	
78	?	Deleted by listwise
84	13	
84	?	Deleted by listwise
85	8	
87	7	
91	7	
92	9	
94	9	
94	11	
96	?	Deleted by listwise
99	7	
105	10	
105	?	Deleted by listwise
106	15	
108	10	
112	?	Deleted by listwise

Imputation (1)

- Statistical based imputation
  - Single Imputation: mengganti nilai yang hilang dengan sebuah nilai estimasi
    - **Mean, Median and Mode** Imputation
      - Mean and median sesuai numerical data, sedangkan mode untuk categorical data
      - Sesuai untuk MCAR
    - Last Observation Carried Forward (LOCF) and Next Observation Carried Forward (NOCF)
  - Multiple Imputation: replacing a missing value with more than one possible value
    - Maximum Likelihood Method: Expectation Maximum Method
    - Matrix Completion Method: Principal Component Analysis (PCA), Probabilistic PCA, and Probabilistic Matrix Factorization
  - Bayesian Approach: missing values are treated as unknown parameters drawn randomly from an appropriate distribution
    - Multivariate imputation by chained equation(MICE)
    - Markov chain Monte Carlo

Imputation (2)

- Machine learning: utilize unsupervised and supervised learning to estimate missing values in datasets, leveraging available information from non-missing data for precise predictions
  - **Regression**: missing values are treated as the dependent variable (Y) and predicted using the other completed independent variables (X)
  - **K-Nearest Neighbour**: selecting the nearest neighbours based on a chosen distance function, K-NN imputes the missing value using the value from the closest neighbor
  - **Clustering**: the information from each cluster can be used to handle missing values
  - **Tree**: Decision Tree and Random Forests
  - **Support Vector Machine**

Imputation (3)

- Neural network: leverage the power of neural networks to learn complex patterns and impute missing values automatically
  - Artificial Neural Network
  - Flow Based
  - Generative Adversarial Networks
  - Diffusion Model

Interpolation (1)

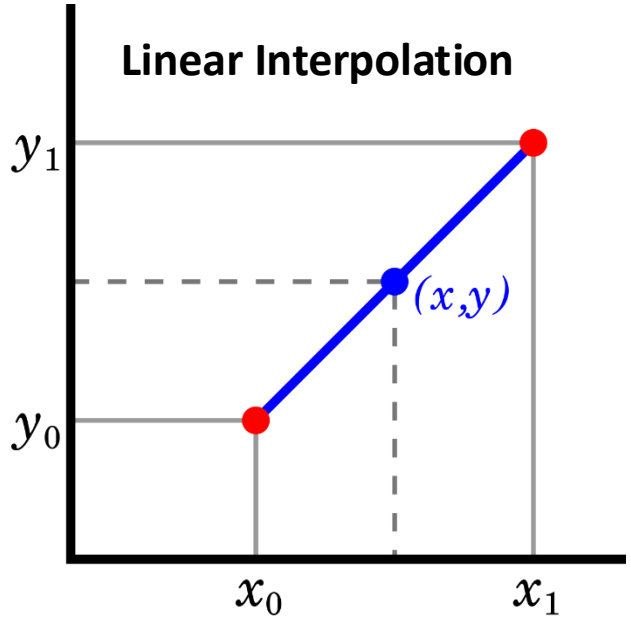
- Piecewise Constant Interpolation
  - Menggantikan nilai yang hilang dengan nilai data terdekat
- Linear Interpolation
  - Estimasi nilai yang hilamg dengan persamaan garis lurus antara dua nilai yang diketahui
    - Persamaan garis dengan trend menaik atau menurun
- Polynomial Interpolation
  - Generalization version of linear interpolation, where the missing values are estimated by using **a curve line of polynomial equation** with **higher degree**
- Spline Interpolation
  - **A curve line of polynomial equation** of low degree (only up to **third degree**)

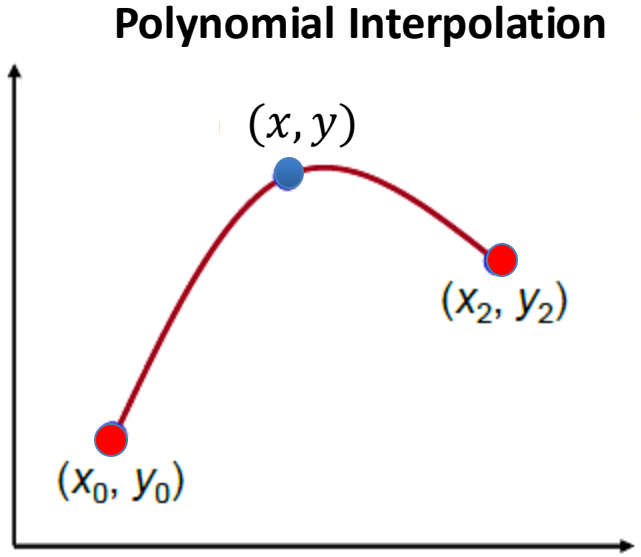
Interpolation (1)

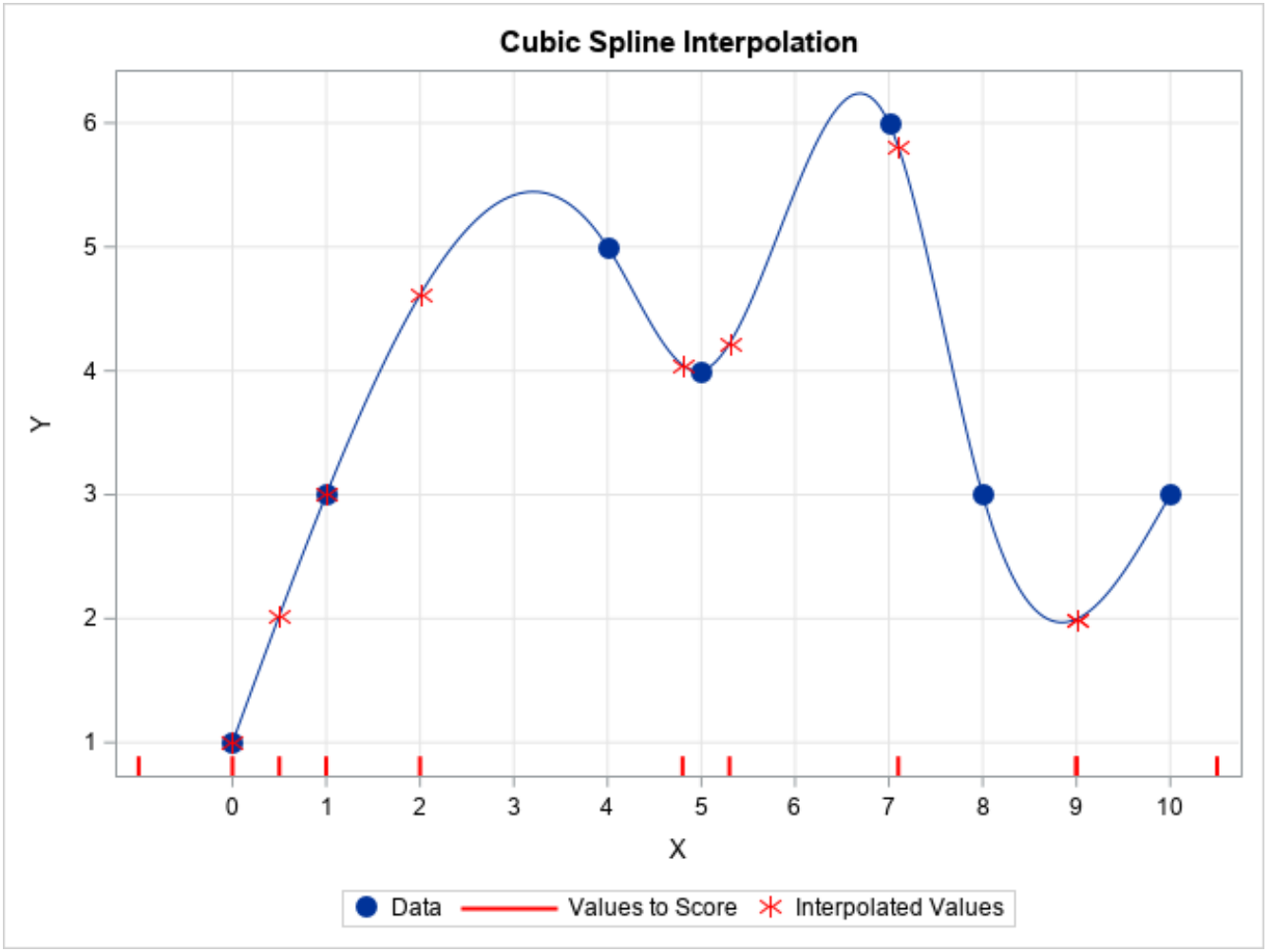
- Piecewise Constant Interpolation
  - Menggantikan nilai yang hilang dengan nilai data terdekat

IQ	Ratings (missing)	Ratings (constant)
84	13	13
84	?	**13**
85	8	8
87	7	7
91	7	7
92	9	9
94	9	9
94	11	11
96	?	**11**
99	7	7
105	10	10
105	?	**10**
106	15	15
108	10	10

Interpolation (2)

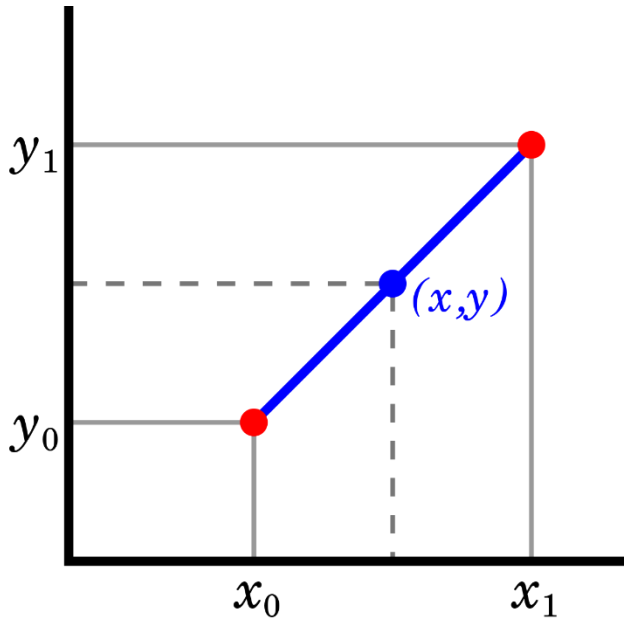






Linear Interpolation

$$y = \frac{y_0(x_1 - x) + y_1(x - x_0)}{(x_1 - x_0)}$$



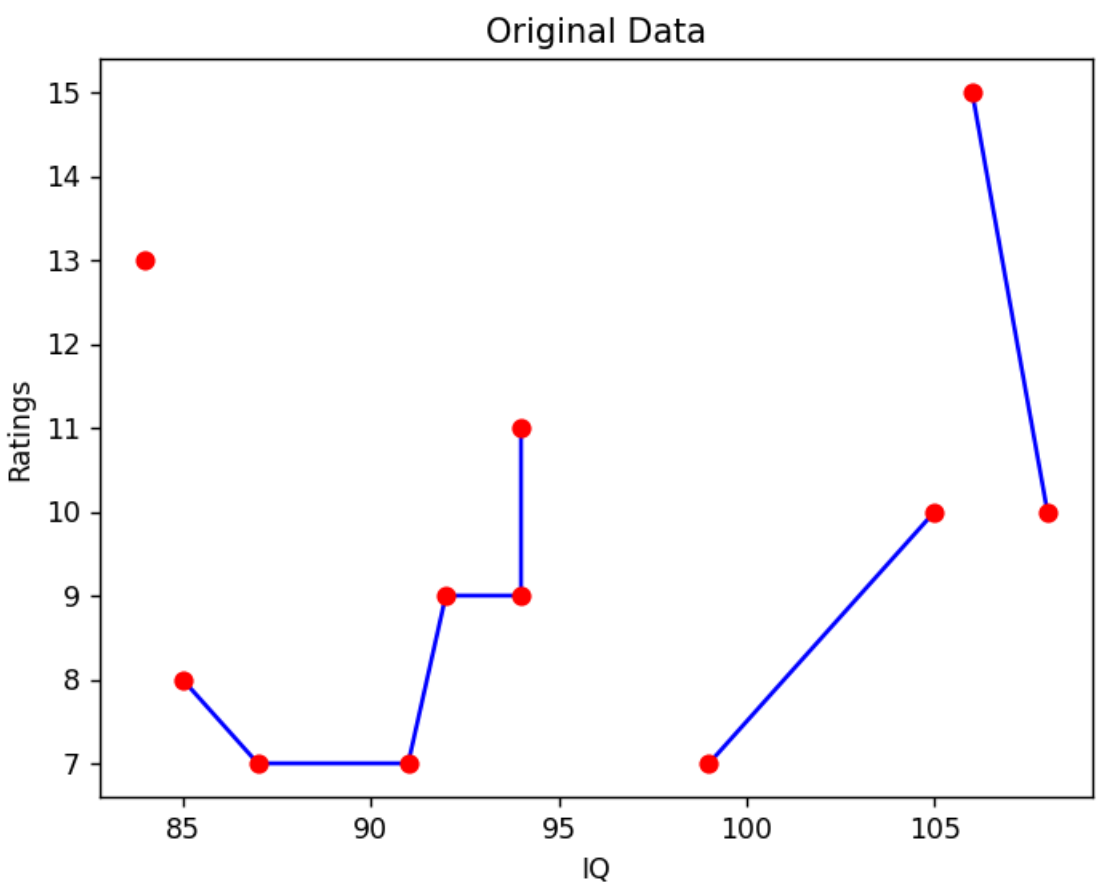
IQ (x)	Ratings (y)	Ratings (Linear)
84	13	13
85	NaN	**10.5**
86	8	8
87	7	7
91	7	7
92	9	9
94	9	9
95	11	11
96	NaN	**10**
99	7	7
105	10	10
106	NaN	**12.5**
107	15	15
108	10	10

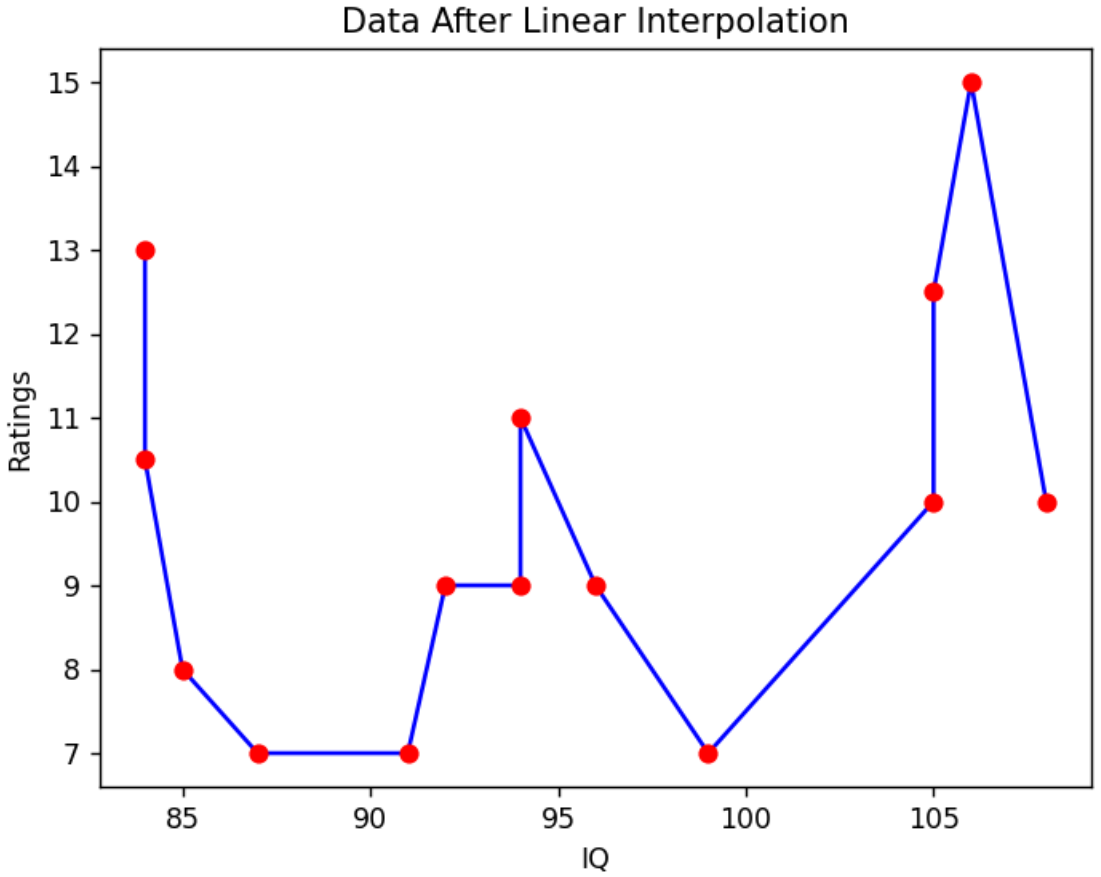
$$y = \frac{13(86 - 85) + 8(85 - 84)}{(86 - 84)} = 10.5$$

$$y = \frac{11(99 - 96) + 7(96 - 95)}{(99 - 95)} = 10$$

$$y = \frac{10(107 - 106) + 15(106 - 105)}{(107 - 105)} = 12.5$$

Linear Interpolation Plot





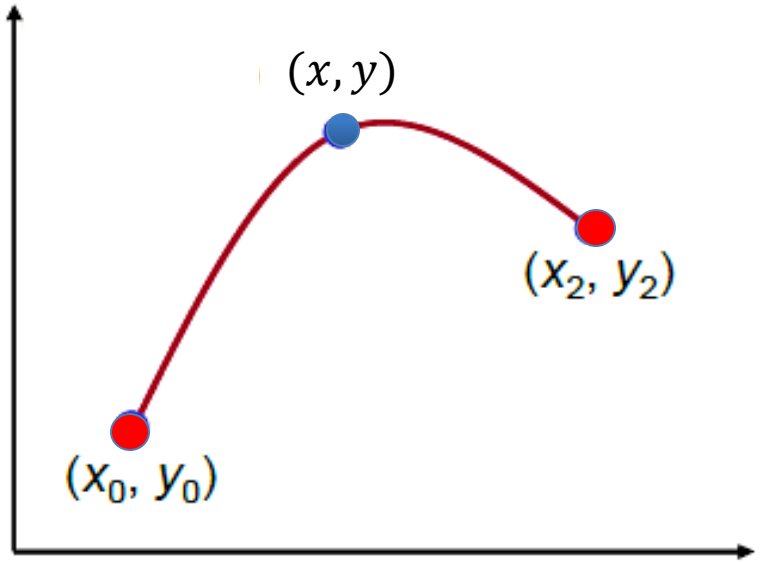
Quadratic Interpolation

$$y = y_0 L_0(x) + y_1 L_1(x) + y_2 L_2(x)$$

$$L_0(x) = \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)}$$

$$L_1(x) = \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)}$$

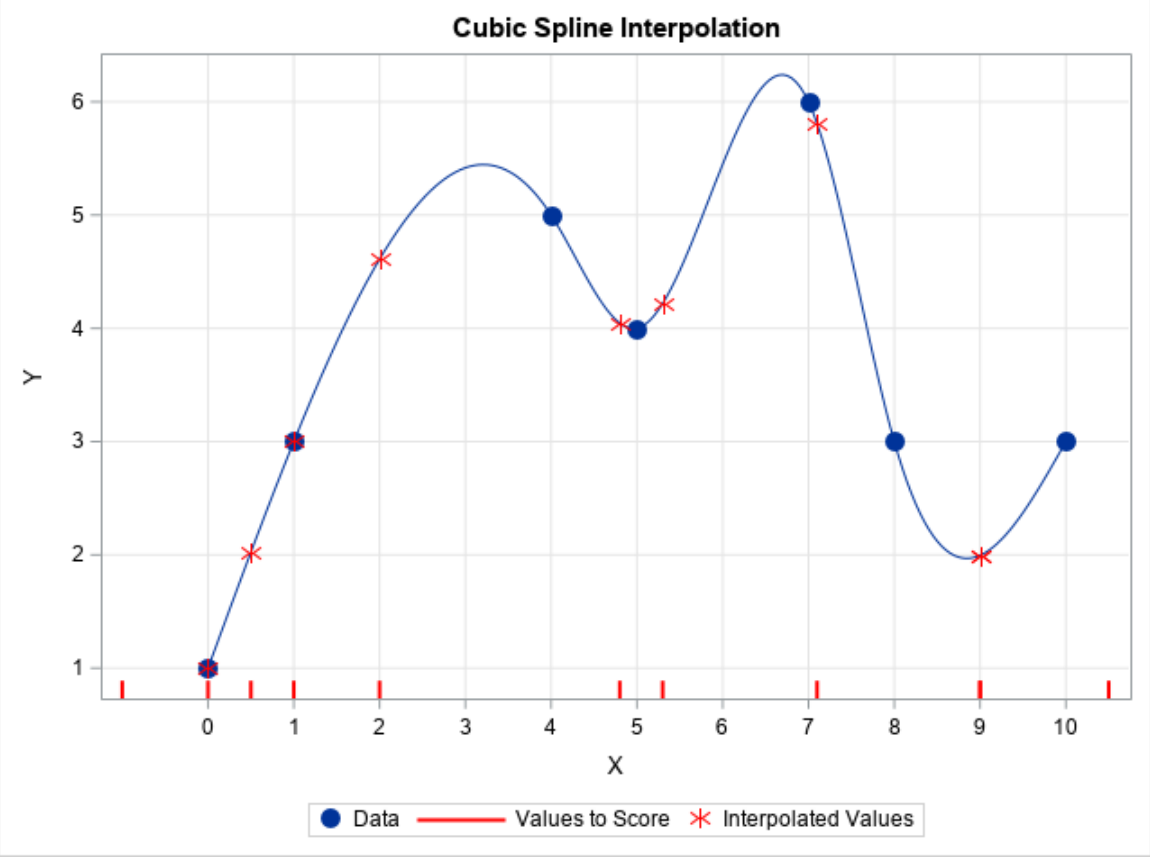
$$L_2(x) = \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)}$$



IQ (x)	Ratings (y)	Ratings (Linear)	Ratings (Quadratic)
84	13	13	13
84	NaN	**10.5**	**9.982980**
85	8	8	8
87	7	7	7
91	7	7	7
92	9	9	9
94	9	9	9
94	11	11	11
96	NaN	**10**	**8.919885**
99	7	7	7
105	10	10	10
105	NaN	**12.5**	**13.943781**
106	15	15	15
108	10	10	10

Polynomial Interpolation

$$f(x) = \begin{cases} -0.1522x^3 + 0.9937x, & \text{if } x \in [0,1], \\ -0.01258x^3 - 0.4189x^2 + 1.4126x - 0.1396, & \text{if } x \in [1,2], \\ 0.1403x^3 - 1.3359x^2 + 3.2467x - 1.3623, & \text{if } x \in [2,3], \\ 0.1579x^3 - 1.4945x^2 + 3.7225x - 1.8381, & \text{if } x \in [3,4], \\ 0.05375x^3 - 0.2450x^2 - 1.2756x + 4.8259, & \text{if } x \in [4,5], \\ -0.1871x^3 + 3.3673x^2 - 19.3370x + 34.9282, & \text{if } x \in [5,6]. \end{cases}$$



IQ (x)	Ratings (y)	Ratings (Linear)	Ratings (Quadratic)	Ratings (Spline)
84	13	13	13	13
84	NaN	10.5	9.982980	9.802086
85	8	8	8	8
87	7	7	7	7
91	7	7	7	7
92	9	9	9	9
94	9	9	9	9
94	11	11	11	11
96	NaN	10	8.919885	9.017973
99	7	7	7	7
105	10	10	10	10
105	NaN	12.5	13.943781	13.945269
106	15	15	15	15
108	10	10	10	10

Interpolation

IQ (x)	Ratings (y)	Ratings (Constant)	Ratings (Linear)	Ratings (Quadratic)	Ratings (Cubic)
84	13	13	13	13	13
84	NaN	**13**	**10.5**	**9.982980**	**9.802086**
85	8	8	8	8	8
87	7	7	7	7	7
91	7	7	7	7	7
92	9	9	9	9	9
94	9	9	9	9	9
94	11	11	11	11	11
96	NaN	**11**	**10**	**8.919885**	**9.017973**
99	7	7	7	7	7
105	10	10	10	10	10
105	NaN	**10**	**12.5**	**13.943781**	**13.945269**
106	15	15	15	15	15
108	10	10	10	10	10

Single Imputation

- each missing value is replaced by:
  - Mean
  - Median
  - Mode

IQ	Ratings (missing)	Ratings (mean)	Ratings (median)	Ratings (mode)
84	13	13	13	13
84	?	**7.57**	**9**	**7**
85	8	8	8	8
87	7	7	7	7
91	7	7	7	7
92	9	9	9	9
94	9	9	9	9
94	11	11	11	11
96	?	**7.57**	**9**	**7**
99	7	7	7	7
105	10	10	10	10
105	?	**7.57**	**9**	**7**
106	15	15	15	15
108	10	10	10	10
Mean	**106/14 = 7.57**			
Median	**9**			
Mode	**7**			