Session-7
Part-A: Data Visualization

1

Importance of Visualization

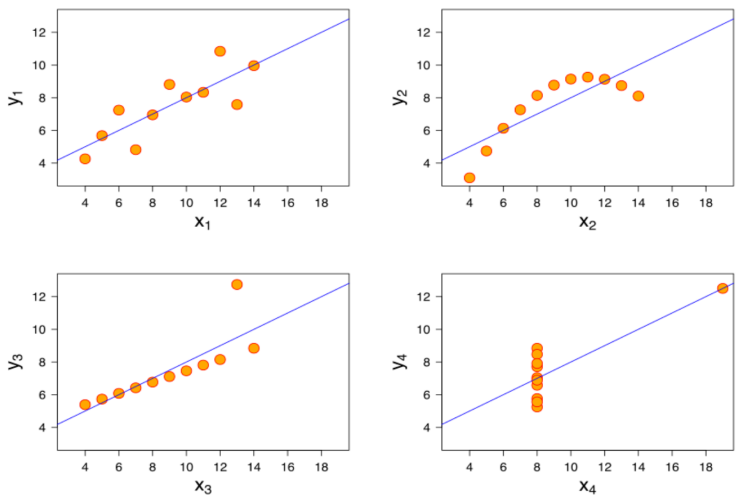
- There are four datasets on the right are known as Anscombe's Quartet
- It was developed by statistician Francis Anscombe
- These four datasets share the same descriptive statistics.

	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	8	5,56
	7	4,82	7	7,26	7	6,42	8	7,91
	5	5,68	5	4,74	5	5,73	8	6,89
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

2

Importance of Visualization

- All four datasets look identical when examined using simple statistics, but vary considerably when we graph them



3

Major Data Visualization Libraries in Python

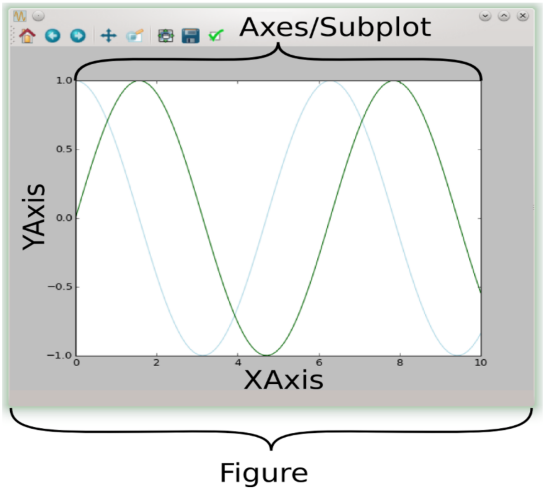
- Matplotlib
- Seaborn
- ggplot
- Altair
- Bokeh
- Plotly
- Folium

4

Matplotlib

Figures
- Overall window/page that everything is drawn on.
- May have multiple independent Figures
- Figures may contain multiple Axes

Axes
- The area where we plot data and any labels associated with it
- Set up Axes with a call to subplot (which places Axes on a regular grid)
- Axes and Subplot are synonymous in most cases
- Each Axes has an XAxis and a YAxis



5

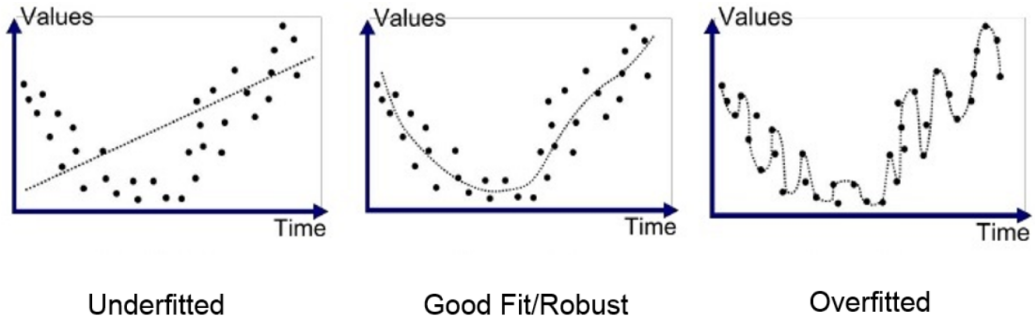
Session-7
Part-B: Model Tuning

6

Overfitting and Underfitting

Overfitting: Good performance on the training data, poor generalization to other data.

Underfitting: Poor performance on the training data and poor generalization to other data.



7

Regularization

A technique that helps appropriately distribute weights among features

- Motivation: Overfitting often caused by overly-complex models capturing idiosyncrasies in training set
- Regularization: Adding penalty score for complexity to cost function

$$cost_{reg} = cost + \frac{\lambda}{2} penalty$$

- Two standard types:
  - L1 regularization: $penalty = \|\vec{w}\|_1 = \sum_{j=1}^{m} |w_j|$
  - L2 regularization: $penalty = \|\vec{w}\|_2^2 = \sum_{j=1}^{m} w_j^2$

8

Regularization – Ridge and Lasso

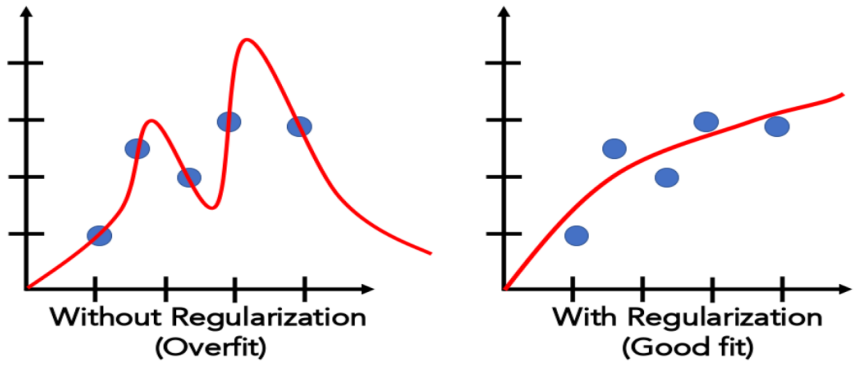
Goal of Regularization is to significantly reduce the variance of the model, without substantial increase in its bias

- **Lasso (L1)**
- Can estimate coefficient of least important features exactly equal to zero when the tuning parameter λ is sufficiently large.
- Therefore, the lasso method also performs variable selection
- **Ridge regression (L2)**
- Shrinks the coefficients for least important predictors, very close to zero.
- But it will never make them exactly zero. In other words, the final model will include all predictors

9

Regularization – Ridge and Lasso (Contd.)

Goal of Regularization is to significantly reduce the variance of the model, without substantial increase in its bias



10


Thank You
ASCEND
towards a sustainable future
KIRON
e-Learning for Future

11