CSE 40647/60647 Data Science (Spring 2018)
Lecture 4: Data Cleaning and Integration

Goals:

- **Understand data quality issues and how to handle them**
  - Describe three types of missing data
  - Describe how to handle missing data, noisy data, inconsistent data, and redundant data
- **Correlation analysis** for handling data redundancy
  - Categorical Variables: Chi-square test
  - Numerical Variables: Covariance analysis

Part I: Quality Issues in Collecting Data:

If you don't want to write down your height or weight…	Sparse data
If you write down you are 12 feet high…	Incomplete data
If in the first lecture you say you are from California and in the third you submit your hometown as Florida…	Noisy data
If in the first lecture you say you are from California and in the third you submit your hometown as C.A. …	Inconsistent data
If you were not in class but your classmate writes down you are from California but actually you are from Florida…	Redundant data

Part II: Missing data types

- Missing Completely at Random (MCAR)
- Missing at Random (MAR)
- Missing Not at Random (MNAR)

Customer	Age	Balance
C1	25	*Missing*
C2	25	100,000
C3	25	*Missing*
C4	60	50,000
C5	60	120,000
C6	60	150,000

Customer	Age	Balance
C1	25	20,000
C2	25	*Missing*
C3	25	15,000
C4	60	50,000
C5	60	*Missing*
C6	60	*Missing*

Customer	Age	Balance
C1	25	20,000
C2	25	100,000
C3	25	*Missing*
C4	60	50,000
C5	60	120,000
C6	60	*Missing*

Part III: Correlation analysis for categorical variables

	Play chess	Not play chess	Sum (row)
Like science fiction	250	200	450
Not like science fiction	50	1000	1050
Sum(col.)	300	1200	1500

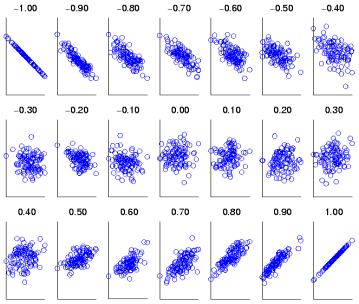
What is the correlation between Play chess and Like science fiction?

Chi-square test:

$$\chi^2 = \sum_{i}^{n} \frac{(O_i - E_i)^2}{E_i}$$

observed

expected



Part IV: Correlation analysis for numerical variables

X1 and X2: (2, 5), (3, 8), (5, 10), (4, 11), (6, 14)

Covariance:

$$\sigma_{12} = E[(X_1 - \mu_1)(X_2 - \mu_2)] = E[X_1 X_2] - \mu_1 \mu_2 = E[X_1 X_2] - E[X_1]E[X_2]$$

Correlation:

$$\rho_{12} = \frac{\sigma_{12}}{\sigma_1 \sigma_2} = \frac{\sigma_{12}}{\sqrt{\sigma_1^2 \sigma_2^2}}$$

	What?	Why?	How to handle?
Incomplete data			
Noisy data			
Inconsistent data			
Redundant data			

Name: **NetID:**

Please write down whatever question you have about this course: