MIS 0855 Spring 2015 – Data Science

Day 19 – Data Cleansing

Min-Seok Pang

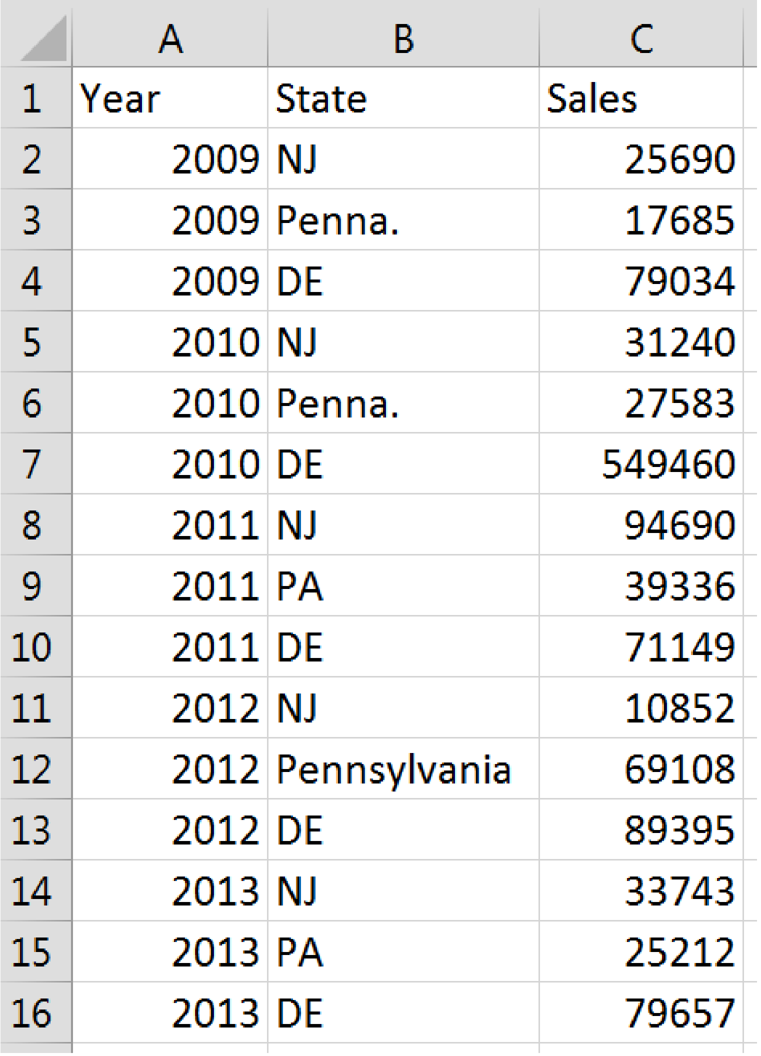
Management Information Systems
Fox School of Business, Temple University
minspang@temple.edu

Feb. 25th, 2015

Fox School of Business
TEMPLE UNIVERSITY

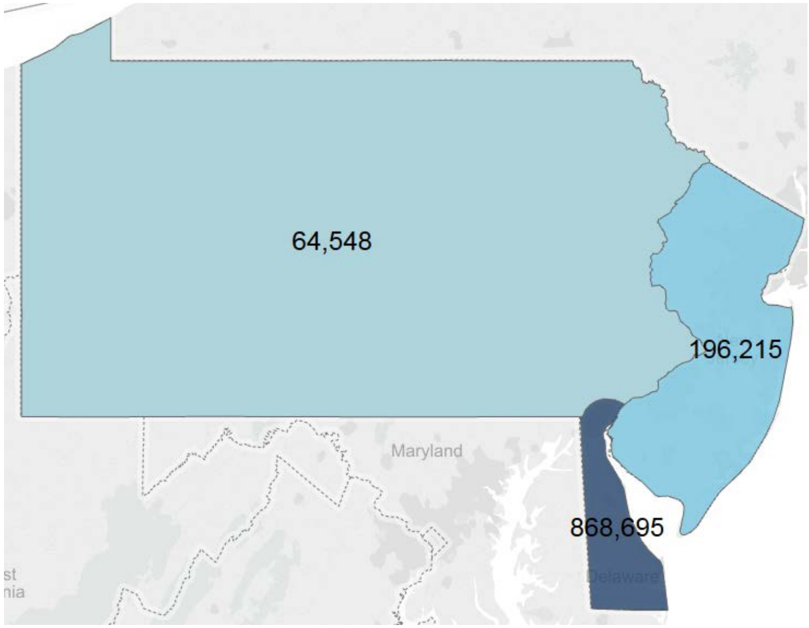
Cleaning Data

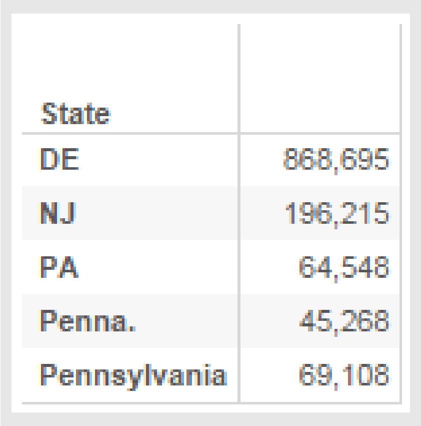
- What are the problems in this dataset?
- What should you do before analysis?



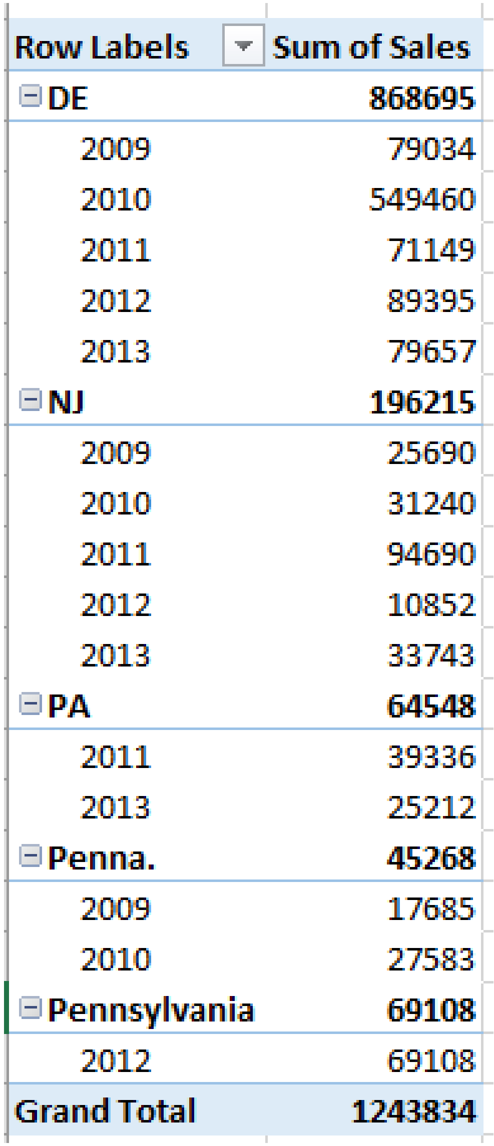
	A	B	C
1	Year	State	Sales
2	2009	NJ	25690
3	2009	Penna.	17685
4	2009	DE	79034
5	2010	NJ	31240
6	2010	Penna.	27583
7	2010	DE	549460
8	2011	NJ	94690
9	2011	PA	39336
10	2011	DE	71149
11	2012	NJ	10852
12	2012	Pennsylvania	69108
13	2012	DE	89395
14	2013	NJ	33743
15	2013	PA	25212
16	2013	DE	79657

Without Data Cleansing





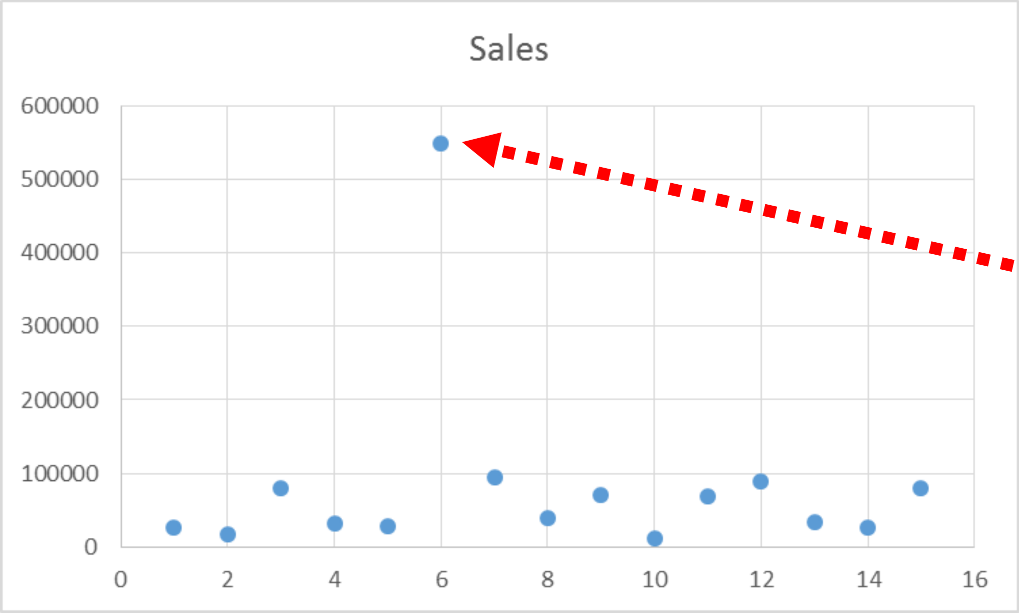
State	
DE	868,695
NJ	196,215
PA	64,548
Penna.	45,268
Pennsylvania	69,108

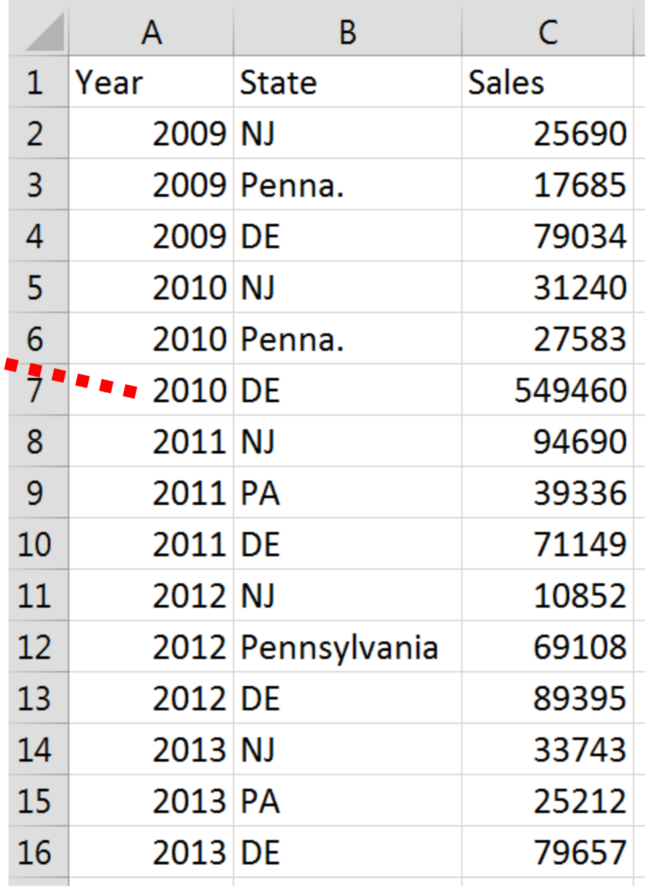


Row Labels	Sum of Sales
⊟DE	868695
2009	79034
2010	549460
2011	71149
2012	89395
2013	79657
⊟NJ	196215
2009	25690
2010	31240
2011	94690
2012	10852
2013	33743
⊟PA	64548
2011	39336
2013	25212
⊟Penna.	45268
2009	17685
2010	27583
⊟Pennsylvania	69108
2012	69108
Grand Total	1243834

- How would you fix this?
- if you have millions of sales records?

Be Careful in Cleaning Data

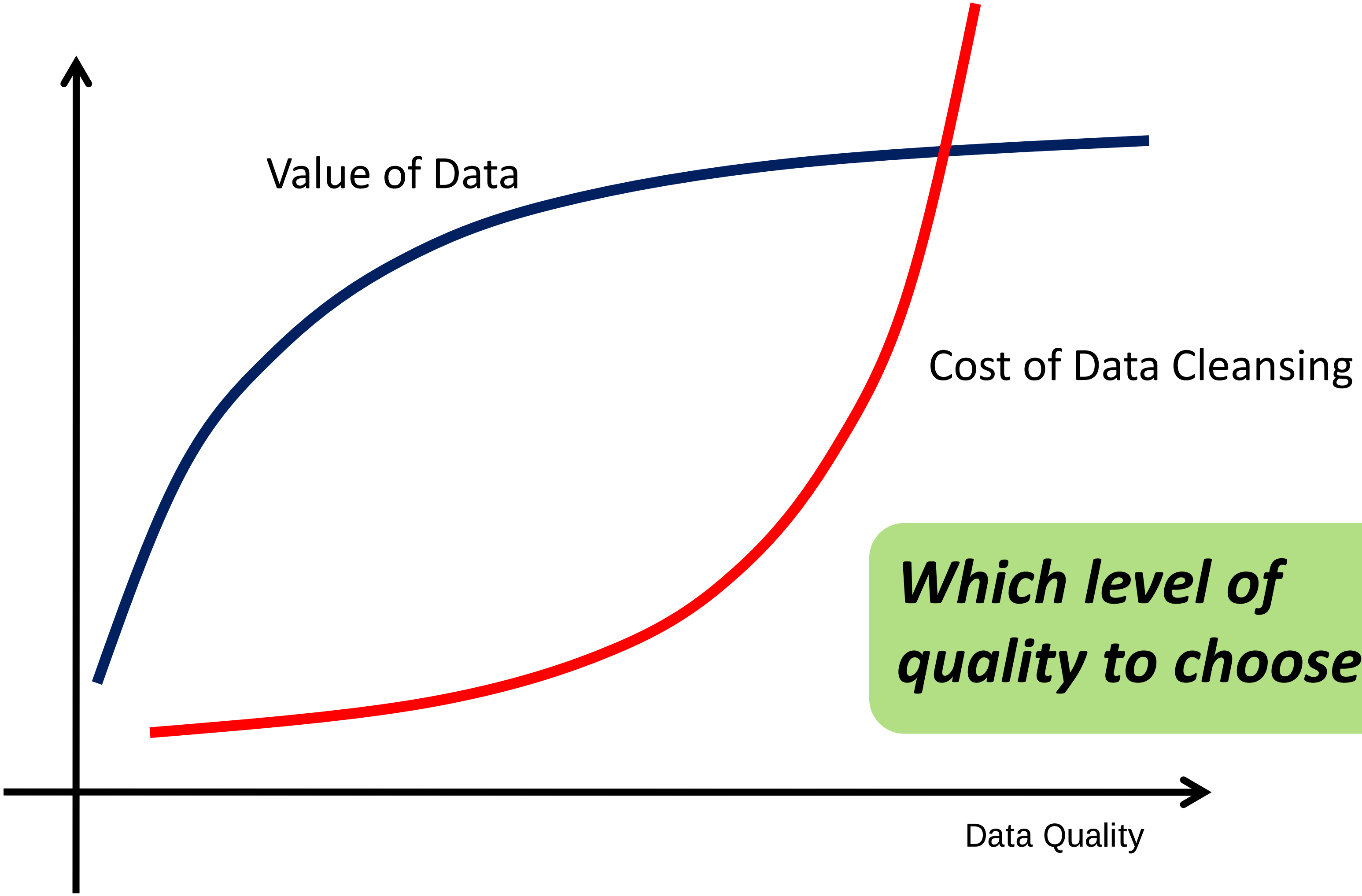




	A	B	C
1	Year	State	Sales
2	2009	NJ	25690
3	2009	Penna.	17685
4	2009	DE	79034
5	2010	NJ	31240
6	2010	Penna.	27583
7	2010	DE	549460
8	2011	NJ	94690
9	2011	PA	39336
10	2011	DE	71149
11	2012	NJ	10852
12	2012	Pennsylvania	69108
13	2012	DE	89395
14	2013	NJ	33743
15	2013	PA	25212
16	2013	DE	79657

- Do we have to remove this?
- If yes, then what?
- Is this really an error or a simply unusual but correct data?

Trade-Off in Data Cleansing



Which level of quality to choose?