Course: INFSCI 2711 Advanced Topics in DBMS

Final Project

Steps to clean the data:

1. Removed records with blank CustomerIDs
2. Removed records weird StockCodes
3. Corrected Description where possible
4. Removed records with where Description cannot be corrected
5. Removed zero Price items
6. Removed “Unspecified” country records
7. Changed Negative quantity records to positive quantity values
8. Separate Datetime into Date and Time columns
9. Calculate subtotal for each record

Column stats for 2009-2010:

No. of records before cleaning: 525461

No. of records after cleaning: 413084

InvoiceNo Range: 489437 – 528618 / C489449 – C538168

StockCode Range: 10002 – 90208 / 10123C – 90214Z & SP1002

Quantity Range: 1 – 19,152

Date Range: Dec 2009 – Dec 2010

Time Range: 7:01 AM – 9:52 PM

UnitPrice Range: $0.08 - $295.00

CustomerID Range: 12346 - 18287

No. of Countries: 36

SubTotal Range: $0.06 - $15,818.40

Column stats for 2010-2011:

No. of records before cleaning: 541910

No. of records after cleaning: 300766

InvoiceNo Range: 536365 – 564272 / C536383 – C564276

StockCode Range: 10002 – 90208 / 10123C – 90214Z

Quantity Range: 1 – 80,995

Date Range: Dec 2010 – Dec 2001

Time Range: 7:35 AM – 8:38 PM

UnitPrice Range: $0.08 - $649.50

CustomerID Range: 12347 - 18287

No. of Countries: 36

SubTotal Range: $0.06 - $168,469.60