DATA PREPROCESSING

WHY DO WE NEED TO PREPROCESS THE DATA?

- Fields that are obsolete or redundant
- Missing values
- Outliers
- Data in a form not suitable for data mining models
- Values not consistent with policy or common sense.

DATA CLEANING

Can You Find Any Problems in This Tiny Data Set?

Customer ID	Zip	Gender	Income	Age	Marital Status	Transaction Amount
1001	10048	M	75000	C	M	5000
1002	J2S7K7	F	−40000	40	W	4000
1003	90210		10000000	45	S	7000
1004	6269	M	50000	0	S	1000
1005	55101	M	99999	30	D	3000

DATA CLEANING - ZIP

Customer ID	Zip
1001	10048
1002	J2S7K7
1003	90210
1004	6269
1005	55101

Standard U.S. zip code = five digits numeral

Customer 1002 zip code of *J2S7K7. (*Actually, this is the zip code of St. Hyancinthe, Quebec, Canada).

Customer 1004? (The zip code is probably *06269, which refers to Storrs, Connecticut, home of the University* of Connecticut*)*

DATA CLEANING - GENDER

Customer ID	Gender
1001	M
1002	F
1003	
1004	M
1005	M

Contains a missing value for customer 1003.

DATA CLEANING - INCOME

Customer ID	Income
1001	75000
1002	−40000
1003	10000000
1004	50000
1005	99999

Customer 1003 is shown as having an income of $10,000,000 per year. Although entirely possible, especially when considering the customer’s zip code (*90210, Beverly Hills), this value of income is nevertheless an outlier, an extreme data value.*

Customer 1004’s reported income of −$40,000 lies beyond the field bounds for income and therefore must be an error.

Customer 1005’s income of $99,999? Perhaps nothing; it may in fact be valid. But if all the other incomes are rounded to the nearest $5000, why the precision with customer 1005? Often, in legacy databases, certain pecified values are meant to be codes for anomalous entries, such as missing values. Perhaps *99999 was coded in an old database to mean missing. Again, we cannot be sure and* should again refer to the “wetware.”

DATA CLEANING – ZIP & INCOME

Customer ID	Zip	Income
1001	10048	75000
1002	J2S7K7	−40000
1003	90210	10000000
1004	6269	50000
1005	55101	99999

Finally, are we clear as to which unit of measure the income variable is measured in? Databases often get merged, sometimes without bothering to check whether such merges are entirely appropriate for all fields. For example, it is quite possible that customer 1002, with the Canadian zip code, has an income measured in Canadian dollars, not U.S. dollars.

DATA CLEANING - AGE

Customer ID	Age
1001	C
1002	40
1003	45
1004	0
1005	30

The *age field has a couple of problems. Although all the other customers have* numerical values for *age, customer 1001's "age" of C probably reflects an earlier categorization* of this man's age into a bin labeled *C. The data mining software will definitely* not like this categorical value in an otherwise numerical field, and we will have to resolve this problem somehow.

How about customer 1004's age of 0? Perhaps there is a *newborn male living in Storrs, Connecticut, who has made a transaction of $1000.*
More likely, the age of this person is probably missing and was coded as 0 to indicate this or some other anomalous condition (e.g., refused to provide the age information).

HANDLING MISSING DATA

- Replace the missing value with some constant, specified by the analyst.
- Replace the missing value with the field mean (for numerical variables) or the mode (for categorical variables).
- Replace the missing values with a value generated at random from the variable distribution observed.

MEAN (RATA-RATA)

Menggambarkan nilai pertengahan dari sekumpulan data

$$\text{Rata} = \sum \frac{X_i}{N}$$

X_i = Kumpulan Data

N = Jumlah Data

REPLACE MISSING VALUE WITH MEAN

Customer ID	Age
1001	C
1002	40
1003	45
1004	20
1005	30

Customer 1001, Nilai C diganti dengan rata-rata Age

Rata = 40 + 45 + 0 + 30 / 4
= 28.75

MODE (MODUS)

Menggambarkan nilai yang paling sering muncul dalam kumpulan data.

Data = $\overset{\downarrow}{1}, 2, \overset{\downarrow}{1}, 4, 3, \overset{\downarrow}{1}, 5, 3, \overset{\downarrow}{1}, 2$

Modus = 1

REPLACE MISSING VALUE WITH MODE

Customer ID	Gender
1001	M
1002	F
1003	N/A
1004	M
1005	M

Customer 1003, Isi Field Gender diganti dengan ‘M’

IDENTIFYING MISCLASSIFICATIONS

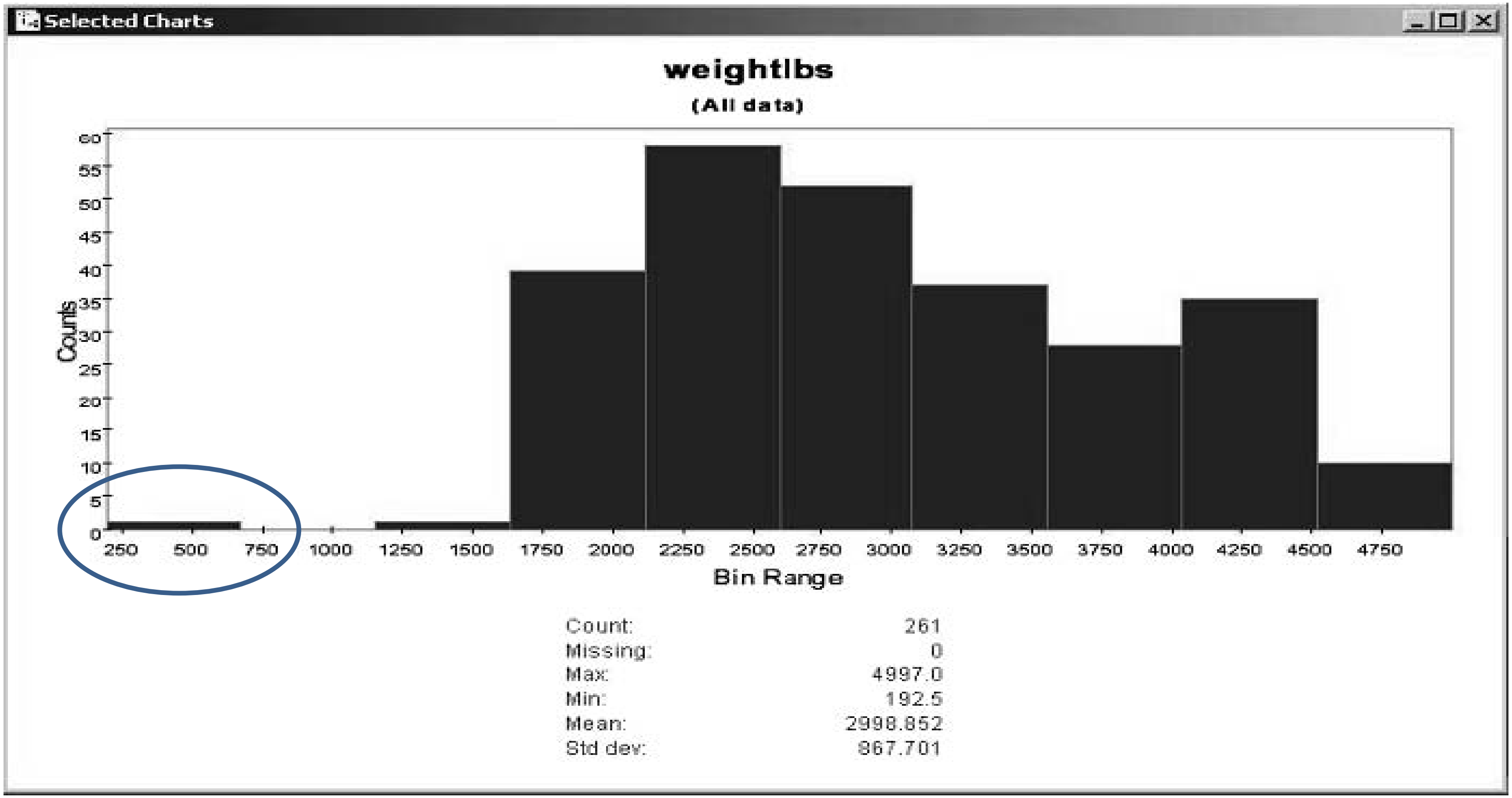
Notice Anything Strange about This Frequency Distribution?

Level Name	Count
USA	1
France	1
US	156
Europe	46
Japan	51

USA and France, have a count of only one automobile each. What is clearly happening here is that two of the records have been classified inconsistently with respect to the origin of manufacture.

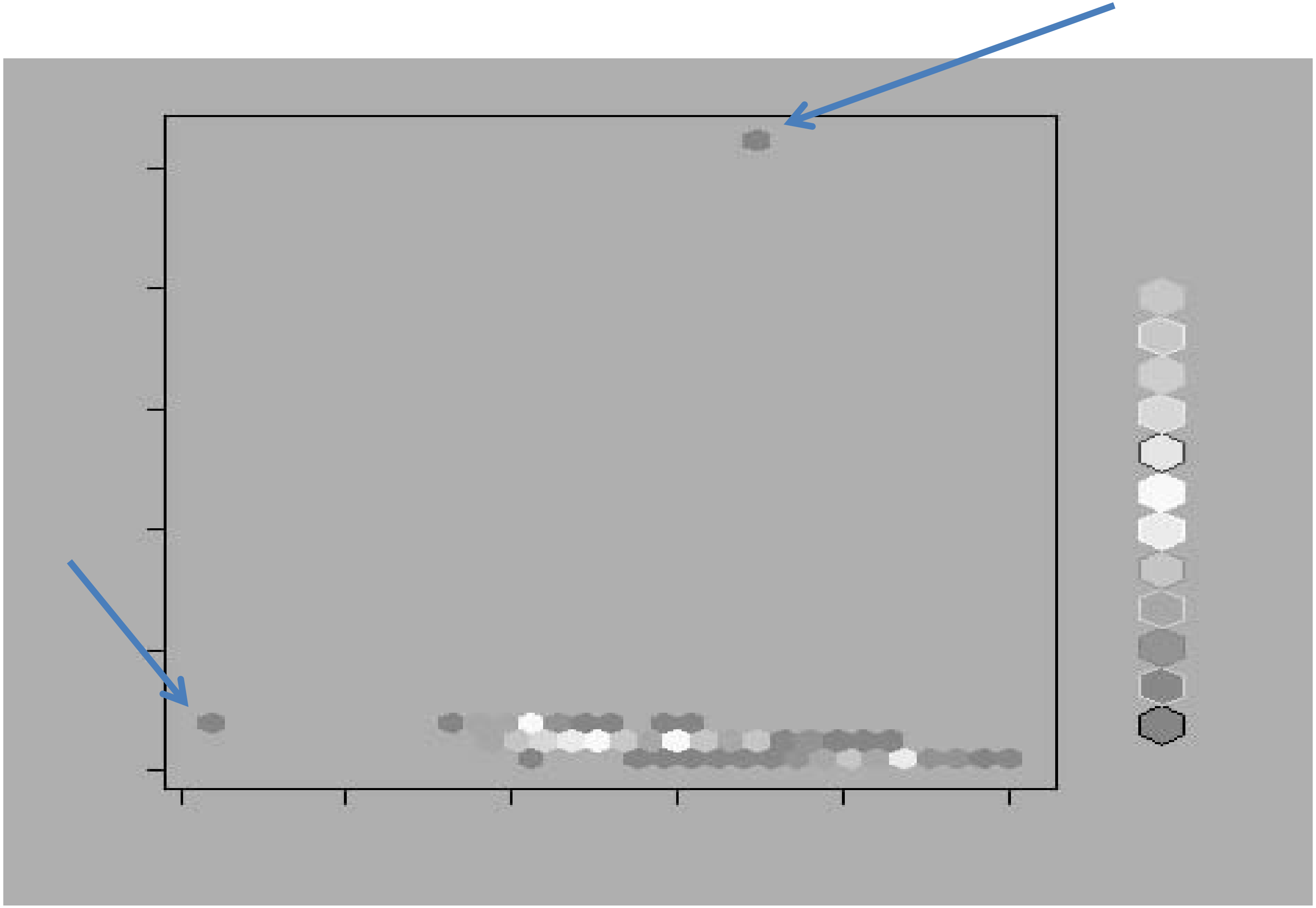
To maintain consistency with the remainder of the data set, the record with origin *USA should have been labeled US, and the record with origin France* should have been labeled *Europe.*

GRAPHICAL METHODS FOR IDENTIFYING OUTLIERS



```
Count:        261
Missing:        0
Max:       4997.0
Min:        192.5
Mean:    2998.852
Std dev:  867.701
```

GRAPHICAL METHODS FOR IDENTIFYING OUTLIERS



DAFTAR PUSTAKA

- Discovering Knowledge in Data (Introduction to Data Mining), Chapter 2, Daniel T. Larose, Wiley, 2004