Applied Statistics

Course Code: MAT1011

Dr. Sukanta Nayak
Department of Mathematics
School of Advanced Sciences
VIT-AP University, Amaravati
Andhra Pradesh

VIT-AP
UNIVERSITY

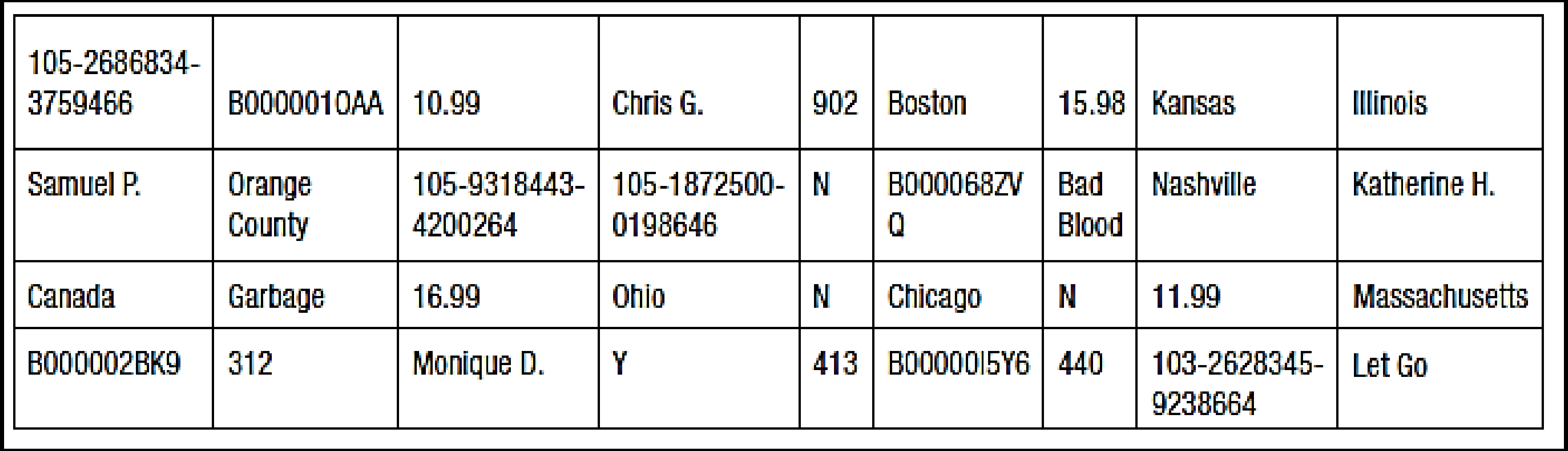
Data

- Collection of information on virtually or physically that are recorded and stored electronically, in vast digital repositories called data warehouses.
- The process of using data, especially of transactional data (data collected for recording the companies’ transactions) to make other decisions and predictions, is sometimes called data mining or predictive analytics. The more general term business analytics (or sometimes simply analytics) describes any use of statistical analysis to drive business decisions from data whether the purpose is predictive or simply descriptive.
- Credit card transactions hold the key to understanding customer behaviour.
- Netflix uses analytics on customer information both to recommend new movies and to adapt the website that customers see to individual tastes.

Data

- To understand better what data are, let’s look at some hypothetical company records that Amazon might collect:

Table 1. An example of data with no context. It’s impossible to say anything about what these values might mean without knowing their context.

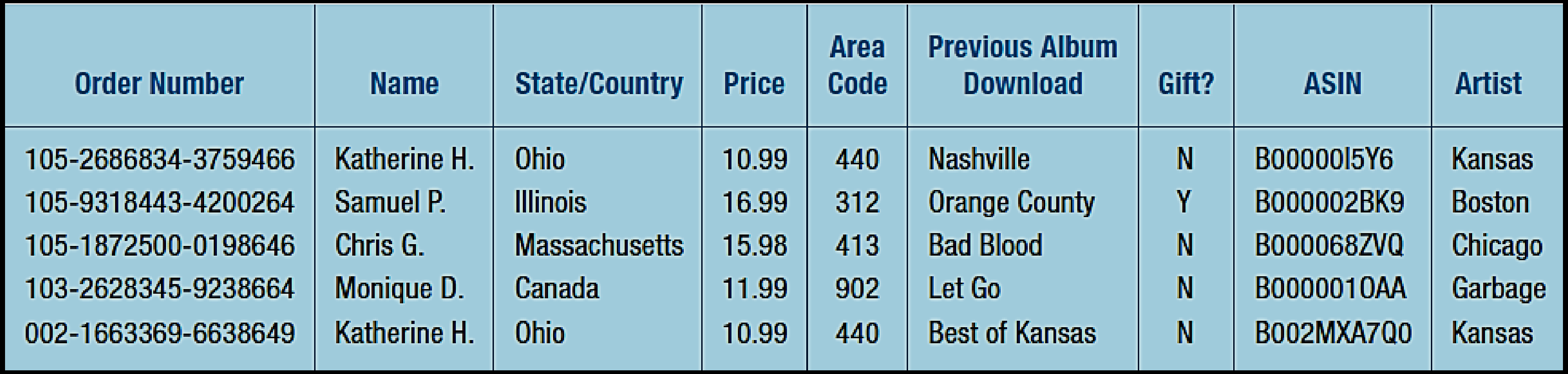


105-2686834-3759466	B0000010AA	10.99	Chris G.	902	Boston	15.98	Kansas	Illinois
Samuel P.	Orange County	105-9318443-4200264	105-1872500-0198646	N	B000068ZVQ	Bad Blood	Nashville	Katherine H.
Canada	Garbage	16.99	Ohio	N	Chicago	N	11.99	Massachusetts
B000002BK9	312	Monique D.	Y	413	B00000I5Y6	440	103-2628345-9238664	Let Go

We can make the meaning clear if we add the context of who the data are about and what was measured and organize the values into a data table such as this one.

Data

Table 2. Example of a data table. The variable names are in the top row. Typically, the Who of the table are found in the leftmost column.

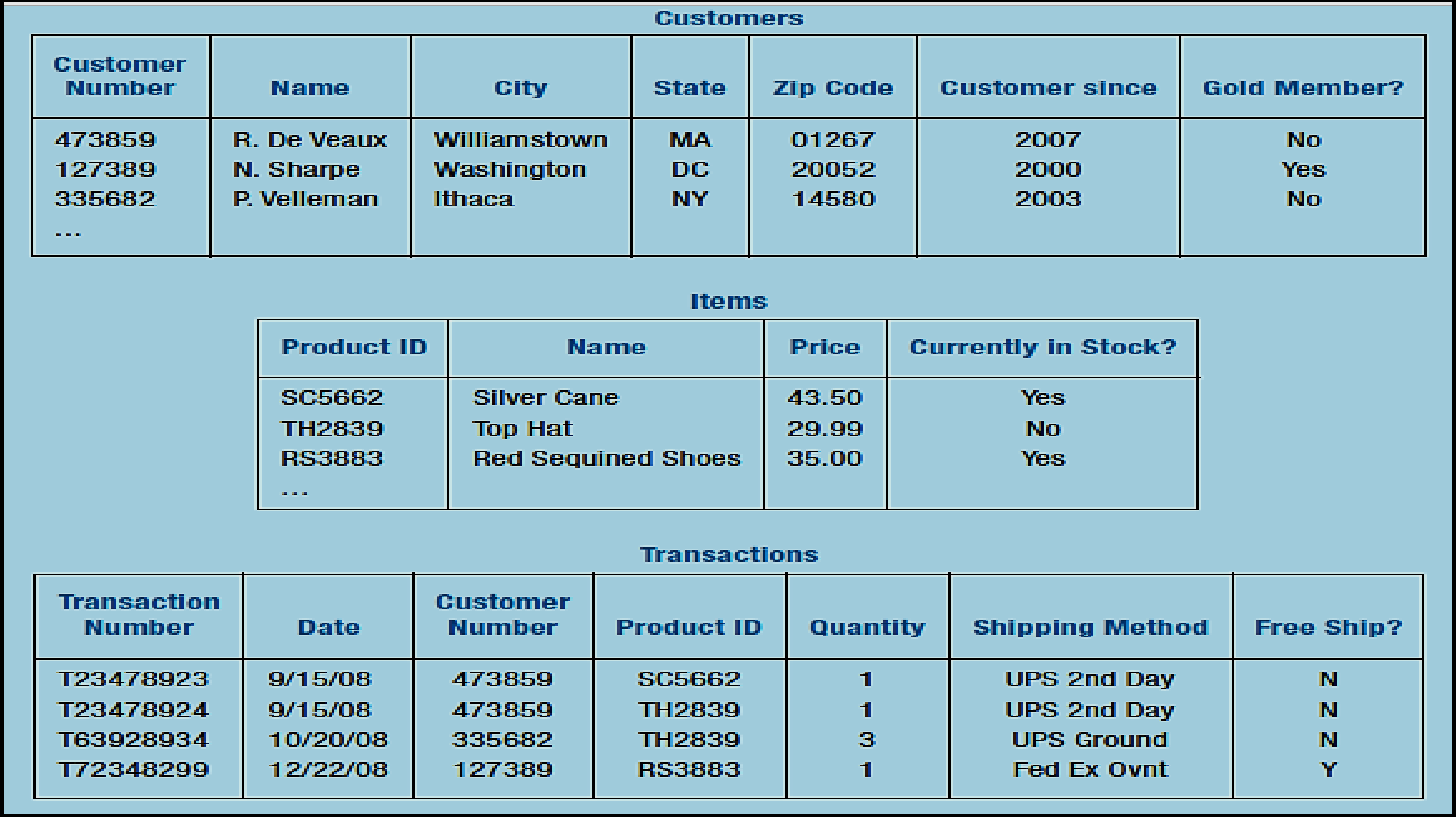


Order Number	Name	State/Country	Price	Area Code	Previous Album Download	Gift?	ASIN	Artist
105-2686834-3759466	Katherine H.	Ohio	10.99	440	Nashville	N	B00000I5Y6	Kansas
105-9318443-4200264	Samuel P.	Illinois	16.99	312	Orange County	Y	B000002BK9	Boston
105-1872500-0198646	Chris G.	Massachusetts	15.98	413	Bad Blood	N	B000068ZVQ	Chicago
103-2628345-9238664	Monique D.	Canada	11.99	902	Let Go	N	B0000010AA	Garbage
002-1663369-6638649	Katherine H.	Ohio	10.99	440	Best of Kansas	N	B002MXA7Q0	Kansas

- In general, the rows of a data table correspond to individual cases about which we’ve recorded some characteristics called variables.
- Individuals who answer a survey are referred to as respondents. People on whom we experiment are subjects or (in an attempt to acknowledge the importance of their role in the experiment) participants, but animals, plants, websites, and other inanimate subjects are often called experimental units. Often we call cases just what they are: for example, customers, economic quarters, or companies. In a database, rows are called records—in this example, purchase records. Perhaps the most generic term is cases.

Data

- Metadata typically contains information about how, when, and where (and possibly why) the data were collected; who each case represents; and the definitions of all the variables.



Customers

Customer Number	Name	City	State	Zip Code	Customer since	Gold Member?
473859	R. De Veaux	Williamstown	MA	01267	2007	No
127389	N. Sharpe	Washington	DC	20052	2000	Yes
335682	P. Velleman	Ithaca	NY	14580	2003	No
...						

Items

Product ID	Name	Price	Currently in Stock?
SC5662	Silver Cane	43.50	Yes
TH2839	Top Hat	29.99	No
RS3883	Red Sequined Shoes	35.00	Yes
...			

Transactions

Transaction Number	Date	Customer Number	Product ID	Quantity	Shipping Method	Free Ship?
T23478923	9/15/08	473859	SC5662	1	UPS 2nd Day	N
T23478924	9/15/08	473859	TH2839	1	UPS 2nd Day	N
T63928934	10/20/08	335682	TH2839	3	UPS Ground	N
T72348299	12/22/08	127389	RS3883	1	Fed Ex Ovnt	Y

Variable Types

- When a variable names categories and answers questions about how cases fall into those categories, we call it a categorical, or qualitative, variable. When a variable has measured numerical values with units and the variable tells us about the quantity of what is measured, we call it a quantitative variable.
- There are exactly as many categories as individuals and only one individual in each category. While it's easy to count the totals for each category, it's not very interesting. This is an identifier variable.
- By contrast, a categorical variable that names categories that don't have order is sometimes called nominal.
- Cross-Sectional and Time Series Data: crosssectional data, where several variables are measured at the same time point.

Data Sources: Where, How, and When

- We must know who, what, and why to analyze data. Without knowing these three, we don’t have enough to start. Of course, we’d always like to know more because the more we know, the more we’ll understand.

What can go wrong?

- Don’t label a variable as categorical or quantitative without thinking about the data and what they represent. The same variable can sometimes take on different roles.
- Don’t assume that a variable is quantitative just because its values are numbers. Categories are often given numerical labels. Don’t let that fool you into thinking they have quantitative meaning. Look at the context.
- Always be skeptical. One reason to analyze data is to discover the truth. Even when you are told a context for the data, it may turn out that the truth is a bit (or even a lot) different. The context colors our interpretation of the data, so those who want to influence what you think may slant the context. A survey that seems to be about all students may in fact report just the opinions of those who visited a fan website. The question that respondents answered may be posed in a way that influences responses.

Terminologies

- Business analytics: The process of using statistical analysis and modeling to drive business decisions.
- Case: A case is an individual about whom or which we have data.
- Cross-sectional data: Data taken from situations that vary over time but measured at a single time instant is said to be a cross-section of the time series.
- Context: The context ideally tells who was measured, what was measured, how the data were collected, where the data were collected, and when and why the study was performed.
- Categorical (or qualitative) variable: A variable that names categories (whether with words or numerals) is called categorical or qualitative.
- Data: Recorded values whether numbers or labels, together with their context.

Terminologies

- Data mining: The process of using a variety of statistical tools to analyze large data bases or data warehouses.
- Data table: An arrangement of data in which each row represents a case and each column represents a variable.
- Data warehouse: A large data base of information collected by a company or other organization usually to record transactions that the organization makes, but also used for analysis via data mining.
- Experimental unit: An individual in a study for which or for whom data values are recorded. Human experimental units are usually called subjects or participants.
- Identifier variable: A categorical variable that records a unique value for each case, used to name or identify it.

Terminologies

- Metadata: Auxiliary information about variables in a database, typically including how, when, and where (and possibly why) the data were collected; who each case represents; and the definitions of all the variables.
- Nominal variable: The term "nominal" can be applied to a variable whose values are used only to name categories.
- Ordinal variable: The term "ordinal" can be applied to a variable whose categorical values possess some kind of order.
- Participant: A human experimental unit. Also called a subject.
- Quantitative variable: A variable in which the numbers are values of measured quantities with units. Record Information about an individual in a database.
- Relational database: A relational database stores and retrieves information. Within the database, information is kept in data tables that can be "related" to each other.

Terminologies

- Respondent: Someone who answers, or responds to, a survey.
- Spreadsheet: A spreadsheet is layout designed for accounting that is often used to store and manage data tables. Excel is a common example of a spreadsheet program.
- Subject: A human experimental unit. Also called a participant.
- Time series: Data measured over time. Usually the time intervals are equally spaced or regularly spaced (e.g., every week, every quarter, or every year).
- Transactional Data: Data collected to record the individual transactions of a company or organization.
- Units: A quantity or amount adopted as a standard of measurement, such as dollars, hours, or grams.
- Variable: A variable holds information about the same characteristic for many cases.

Summary

- Identify whether a variable is being used as categorical or quantitative.
- Categorical variables identify a category for each case. Usually we think about the counts of cases that fall in each category. (An exception is an identifier variable that just names each case.)
- Quantitative variables record measurements or amounts of something; they must have units.
- Sometimes we may treat the same variable as categorical or quantitative depending on what we want to learn from it, which means some variables can’t be pigeonholed as one type or the other.
- Consider the source of your data and the reasons the data were collected. That can help you understand what you might be able to learn from the data.

Summary

- Understand that data are values, whether numerical or labels, together with their context.
- who, what, why, where, when (and how)—the W's—help nail down the context of the data.
- We must know who, what, and why to be able to say anything useful based on the data. The who are the cases. The what are the variables. A variable gives information about each of the cases. The why helps us decide which way to treat the variables.
- Stop and identify the W's whenever you have data, and be sure you can identify the cases and the variables.

Task

CATEGORY	NAME
HOME & KITCHEN	
MOBILE & ACCESSORIES	
PERSONAL HEALTH, GROOMING & WELLNESS	
ELECTRONICS & ACCESSORIES	
COMPUTER & ACCESSORIES	
TV & APPLIANCES	
WOMEN’S FASHION	
MEN’S FASHION	
KID’S FASHION	
SPORTS & FITNESS	

References

1. Norean R. Sharpe, Richard D. De Veaux, Paul F. Velleman, Business Statistics, Fourth Edition, Pearson Education. 2019.
2. R.E. Walpole, R.H. Mayers, S.L. Mayers and K.Ye, Probability and Statistics for engineers and scientists, 9th Edition, Pearson Education, 2018.
3. S. C. Gupta and V. K. Kapoor, Fundamentals of Applied Statistics, S. Chand, 2006.
4. Douglas C. Montgomery, George C. Runger, Applied Statistics and Probability for Engineers, Sixth Edition, John Wiley and Sons, 2014.
5. Miller & Freund’s, Probability and statistics for engineers, 8th edition, Pearson publication, 2018.

THE END