BUSINESS ANALYTICS

DR. BRENDA MULLALLY

DATA EVERYWHERE

- TECHNOLOGY HAS MADE IT POSSIBLE TO COLLECT AND STORE HUGE AMOUNTS OF DATA.
  - RETAILERS, CREDIT AGENCIES, INVESTMENT COMPANIES, GOVERNMENT AGENCIES,
- IT IS DIFFICULT FOR BUSINESSES TO MAKE SENSE OF ALL OF THE DATA COLLECTED.
- MANY MORE PEOPLE NOW HAVE THE POWER TO ANALYSE DATA AND MAKE DECISIONS ON THE BASIS OF QUANTITATIVE ANALYSIS.
- QUANTITATIVE ANALYSIS IS NOW CONDUCTED BY PEOPLE OTHER THAN THOSE THAT TRADITIONALLY HAD DONE THE NUMBER CRUNCHING.
- MOST EMPLOYEES NOW HAVE ACCESS TO SOFTWARE TO ANALYSE DATA, PARTICULARLY SPREADSHEET AND DATABASE SOFTWARE.
- QUANTITATIVE ANALYSIS IS NOW AN INTEGRAL PART OF THESE PEOPLE'S JOB.

DATA EXPLORATION

- BASIC DATA SUMMARIES AND VISUALISATIONS:
  - SUMMARY STATISTICS
  - FREQUENCY TABLES
  - HISTOGRAMS
  - BOXPLOTS
  - SCATTERPLOTS
  - CORRELATION TABLES
  - CROSS-TABULATIONS

DATA EXPLORATION

- TYPICAL EMPLOYEES TODAY NOT JUST THE MANAGERS AND TECHNICAL SPECIALISTS HAVE A WEALTH OF EASY-TO-USE TOOLS AT THEIR DISPOSAL, AND IT IS FREQUENTLY UP TO THEM TO SUMMARIZE DATA IN A WAY THAT IS BOTH MEANINGFUL AND USEFUL TO THEIR CONSTITUENTS: PEOPLE WITHIN THEIR COMPANY, THEIR COMPANY'S SUPPLIERS, AND THEIR COMPANY'S CUSTOMERS. IT TAKES SOME TRAINING AND PRACTICE TO DO THIS EFFECTIVELY.

DATA EXPLORATION

- DATA ANALYSIS IN THE REAL WORLD IS NEVER DONE IN A VACUUM. IT IS DONE TO SOLVE A PROBLEM. TYPICALLY, THERE ARE FOUR STEPS THAT ARE FOLLOWED, WHETHER THE CONTEXT IS BUSINESS, MEDICAL SCIENCE, OR ANY OTHER FIELD.

1. RECOGNISE A PROBLEM THAT NEEDS SOLVING
2. GATHER DATA TO HELP UNDERSTAND AND THEN SOLVE THE PROBLEM.
3. ANALYSE THE DATA
4. ACT ON THE ANALYSIS BY CHANGING POLICIES, UNDERTAKING INITIATIVES, PUBLISHING RECORDS ETC.

DATA EXPLORATION

- POPULATIONS AND SAMPLES
  - POPULATION INCLUDES ALL OF THE ENTITIES OF INTEREST: PEOPLE, HOUSEHOLDS, MACHINES, OR WHATEVER.
  - SAMPLE IS A SUBSET OF A POPULATION, OFTEN RANDOMLY CHOSEN AND PREFERABLY REPRESENTATIVE OF THE POPULATION AS A WHOLE.
- IT IS VERY IMPORTANT THAT THE SAMPLE IS REPRESENTATIVE OF THE POPULATION. THIS MEANS THAT ANY OBSERVED CHARACTERISTICS OF THE SAMPLE CAN BE GENERALISED TO THE POPULATION AS A WHOLE.

DATA EXPLORATION

CROSS-SECTIONAL DATA

- DATA GATHERED AT A SINGLE POINT IN TIME FROM DIFFERENT INDIVIDUALS OR GROUPS.
- STOCK INVENTORY, HUMAN OPINION, GRADES FOR A MODULE IN A SEMESTER.

TIME SERIES DATA

- DATA GATHERED AT USUALLY DISCRETE AND EQUALLY SPACED TIME INTERVALS.
- DAILY CLOSING PRICE OF STOCK, DAILY TEMPERATURE, STAFF NUMBERS EACH MONTH, WEEKLY SALES, STUDENT REGISTRATION ON A COURSE EACH YEAR.

DATA EXPLORATION

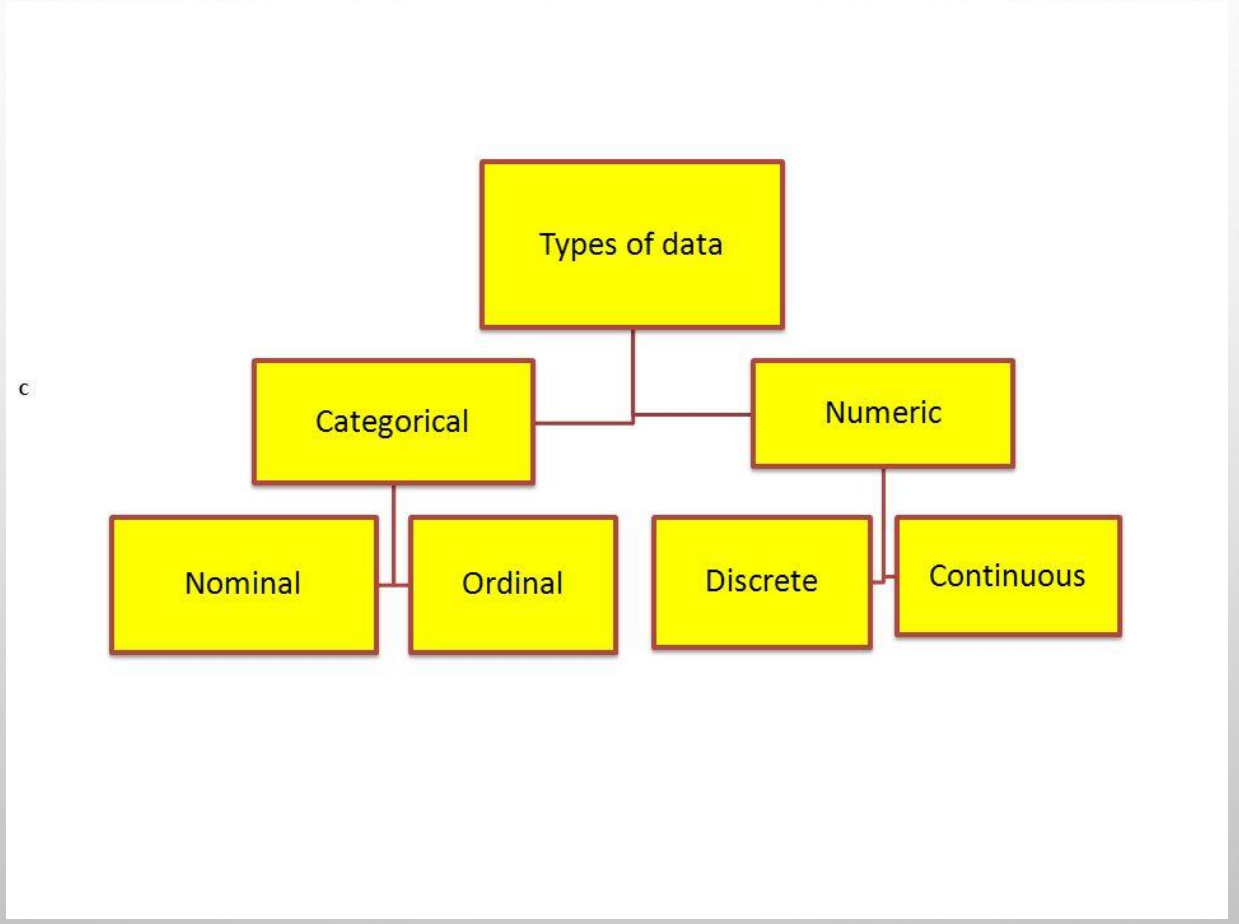
- DATA SETS, VARIABLES, AND OBSERVATIONS
  - DATA SET: A RECTANGULAR ARRAY OF DATA WHERE COLUMNS CONTAIN VARIABLES, SUCH AS HEIGHT, GENDER, AND INCOME.
  - EACH ROW CONTAINS AN OBSERVATION.
  - EACH OBSERVATION CONTAINS THE ATTRIBUTES OF A PARTICULAR MEMBER OF A POPULATION: A PERSON, A COMPANY, A CITY, A MACHINE…
  - A VARIABLE (COLUMN) IS OFTEN CALLED A FIELD OR AN ATTRIBUTE.
  - AN OBSERVATION (ROW) IS OFTEN CALLED A CASE OR A RECORD.

EXAMPLE 2.1: QUESTIONNAIRE DATA.XLSX

- **OBJECTIVE:** TO ILLUSTRATE VARIABLES AND OBSERVATIONS IN A TYPICAL DATA SET.
- **SOLUTION:** DATA SET INCLUDES OBSERVATIONS ON 30 PEOPLE WHO RESPONDED TO A QUESTIONNAIRE ON THE PRESIDENT'S ENVIRONMENTAL POLICIES.
- VARIABLES INCLUDE: AGE, GENDER, STATE, CHILDREN, SALARY, OPINION.
- INCLUDES A ROW THAT LISTS VARIABLE NAMES.
- INCLUDES A COLUMN THAT SHOWS AN INDEX OF THE OBSERVATION.

	A	B	C	D	E	F	G
1	Person	Age	Gender	State	Children	Salary	Opinion
2	1	35	Male	Minnesota	1	$65,400	5
3	2	61	Female	Texas	2	$62,000	1
4	3	35	Male	Ohio	0	$63,200	3
5	4	37	Male	Florida	2	$52,000	5
6	5	32	Female	California	3	$81,400	1
7	6	33	Female	New York	3	$46,300	5
28	27	27	Male	Illinois	3	$45,400	2
29	28	63	Male	Michigan	2	$53,900	1
30	29	52	Male	California	1	$44,100	3
31	30	48	Female	New York	2	$31,000	4

DATA EXPLORATION



DATA EXPLORATION

- DATA TYPES
  - NUMERICAL AND CATEGORICAL DATA
  - DO YOU WANT TO DO ARITHMETIC ON THE DATA?
  - CAN YOU AVERAGE DAYS OF THE WEEK OR GENDER?
  - WHAT ABOUT A VARIABLE THAT HAS 1, 2, 3, 4, OR 5 AS ITS VALUE?
  - ORDINAL: A NATURAL ORDERING TO CATEGORIES.
  - NOMINAL: NO NATURAL ORDER TO CATEGORIES.
  - ALL CATEGORICAL VARIABLES CAN BE ENCODED WITH NUMBERS BUT NOT ALL ARE, IT IS PERSONAL CHOICE.
  - DUMMY VARIABLE

DATA EXPLORATION

- A NUMERICAL VARIABLE IS **DISCRETE** IF IT RESULTS FROM A COUNT, SUCH AS THE NUMBER OF CHILDREN.
- A **CONTINUOUS** VARIABLE IS THE RESULT OF AN ESSENTIALLY CONTINUOUS MEASUREMENT, SUCH AS WEIGHT OR HEIGHT.
- DATA SET:
- **CROSS-SECTIONAL** DATA ARE DATA ON A CROSS SECTION OF A POPULATION AT A DISTINCT POINT IN TIME.
- **TIME SERIES** DATA ARE DATA COLLECTED OVER TIME.

DATA EXPLORATION

- DATA TYPES
  - SOMETIMES A NUMBER VARIABLE IS CODED USING A CATEGORY.
  - BINNING (DISCRETISING)

ENVIRONMENTAL DATA
USING A DIFFERENT CODING

	A	B	C	D	E	F	G
1	**Person**	**Age**	**Gender**	**State**	**Children**	**Salary**	**Opinion**
2	1	Middle-aged	1	Minnesota	1	$65,400	Strongly agree
3	2	Elderly	0	Texas	2	$62,000	Strongly disagree
4	3	Middle-aged	1	Ohio	0	$63,200	Neutral
5	4	Middle-aged	1	Florida	2	$52,000	Strongly agree
6	5	Young	0	California	3	$81,400	Strongly disagree
7	6	Young	0	New York	3	$46,300	Strongly agree
8	7	Elderly	0	Minnesota	2	$49,600	Strongly disagree
9	8	Middle-aged	1	New York	1	$45,900	Strongly agree
10	9	Middle-aged	1	Texas	3	$47,700	Agree
11	10	Young	0	Texas	1	$59,900	Agree
12	11	Middle-aged	1	New York	1	$48,100	Agree
13	12	Middle-aged	0	Virginia	0	$58,100	Neutral
14	13	Middle-aged	0	Illinois	2	$56,000	Strongly disagree
15	14	Middle-aged	0	Virginia	2	$53,400	Strongly disagree
16	15	Middle-aged	0	New York	2	$39,000	Disagree
17	16	Middle-aged	1	Michigan	1	$61,500	Disagree
18	17	Middle-aged	1	Ohio	0	$37,700	Strongly disagree
19	18	Middle-aged	0	Michigan	2	$36,700	Agree
28	27	Young	1	Illinois	3	$45,400	Disagree
29	28	Elderly	1	Michigan	2	$53,900	Strongly disagree
30	29	Middle-aged	1	California	1	$44,100	Neutral
31	30	Middle-aged	0	New York	2	$31,000	Agree

Note the formulas in columns B, C, and G that generate this recoded data. The formulas in columns B and G are based on the lookup tables below.

Age lookup table (range name AgeLookup)

0	Young
35	Middle-aged
60	Elderly

Opinion lookup table (range name OpinionLookup)

1	Strongly disagree
2	Disagree
3	Neutral
4	Agree
5	Strongly agree

DATA EXPLORATION

- MOST REAL DATA SETS HAVE GAPS IN THE DATA.
- THERE ARE TWO ISSUES: HOW TO DETECT THESE **MISSING VALUES** AND WHAT TO DO ABOUT THEM.
- THE MORE IMPORTANT ISSUE IS WHAT TO DO ABOUT THEM:
  - ONE OPTION IS TO SIMPLY IGNORE THEM. THEN YOU WILL HAVE TO BE AWARE OF HOW THE SOFTWARE DEALS WITH MISSING VALUES.
  - ANOTHER OPTION IS TO FILL IN MISSING VALUES WITH THE AVERAGE OF NON MISSING VALUES, BUT THIS ISN'T USUALLY A VERY GOOD OPTION.
  - A THIRD OPTION IS TO EXAMINE THE NONMISSING VALUES IN THE *ROW* OF A MISSING VALUE; THESE VALUES MIGHT PROVIDE CLUES ON WHAT THE MISSING VALUE SHOULD BE.

EXCEL TABLES FOR FILTERING, SORTING, AND SUMMARIZING

- TABLES ARE A TOOL INTRODUCED IN EXCEL 2007.
- YOU NOW HAVE THE ABILITY TO DESIGNATE A RECTANGULAR DATA SET AS A TABLE AND THEN EMPLOY A NUMBER OF POWERFUL TOOLS FOR ANALYZING TABLES.
- THESE TOOLS INCLUDE:
  - FILTERING
  - SORTING
  - SUMMARIZING

EXAMPLE 2.7:CATALOG MARKETING.XLSX

- **OBJECTIVE**: TO ILLUSTRATE EXCEL TABLES FOR ANALYZING THE HYTEX DATA.
- **SOLUTION**: DATA SET CONTAINS DATA ON 1000 CUSTOMERS OF HYTEX, A FICTIONAL DIRECT MARKETING COMPANY.
- DESIGNATE THE DATA SET AS A TABLE BY SELECTING ANY CELL IN THE DATA SET AND CLICKING THE TABLE BUTTON ON THE INSERT RIBBON.
- USE THE DROPDOWN ARROWS NEXT TO THE VARIABLE NAMES TO FILTER IN MANY DIFFERENT WAYS.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Person	Age	Gender	Own Home	Married	Close	Salary	Children	History	Catalogs	Region	State	City	First Purchase	Amount Spent
2	1	1	0	0	0	1	$16,400	1	1	12	South	Florida	Orlando	10/23/2008	$218
3	2	2	0	1	1	0	$108,100	3	3	18	Midwest	Illinois	Chicago	5/25/2006	$2,632
4	3	2	1	1	1	1	$97,300	1	NA	12	South	Florida	Orlando	8/18/2012	$3,048
5	4	3	1	1	1	1	$26,800	0	1	12	East	Ohio	Cleveland	12/26/2009	$435
6	5	1	1	0	0	1	$11,200	0	NA	6	Midwest	Illinois	Chicago	8/4/2012	$106
7	6	2	0	0	0	1	$42,800	0	2	12	West	Arizona	Phoenix	3/4/2010	$759
8	7	2	0	0	0	1	$34,700	0	NA	18	Midwest	Kansas	Kansas City	6/11/2012	$1,615
9	8	3	0	1	1	0	$80,000	0	3	6	West	California	San Francisco	8/17/2006	$1,985
10	9	2	1	1	0	1	$60,300	0	NA	24	Midwest	Illinois	Chicago	5/29/2012	$2,091
11	10	3	1	1	1	0	$62,300	0	3	24	South	Florida	Orlando	6/9/2008	$2,644

CATALOG MARKETING.XLSX

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Person	Age	Gender	Own Home	Married	Close	Salary	Children	History	Catalogs	Region	State	City	First Purchase	Amount Spent
2	1	1	0	0	0	1	$16,400	1	1	12	South	Florida	Orlando	10/23/2008	$218
3	2	2	0	1	1	0	$108,100	3	3	18	Midwest	Illinois	Chicago	5/25/2006	$2,632
4	3	2	1	1	1	1	$97,300	1	NA	12	South	Florida	Orlando	8/18/2012	$3,048
5	4	3	1	1	1	1	$26,800	0	1	12	East	Ohio	Cleveland	12/26/2009	$435
6	5	1	1	0	0	1	$11,200	0	NA	6	Midwest	Illinois	Chicago	8/4/2012	$106
7	6	2	0	0	0	1	$42,800	0	2	12	West	Arizona	Phoenix	3/4/2010	$759
8	7	2	0	0	0	1	$34,700	0	NA	18	Midwest	Kansas	Kansas City	6/11/2012	$1,615
9	8	3	0	1	1	0	$80,000	0	3	6	West	California	San Francisco	8/17/2006	$1,985
10	9	2	1	1	0	1	$60,300	0	NA	24	Midwest	Illinois	Chicago	5/29/2012	$2,091

FILTERING

- FINDING RECORDS THAT MATCH PARTICULAR CRITERIA IS CALLED *FILTERING*.
- ONE WAY TO FILTER IS TO CREATE AN EXCEL TABLE, WHICH AUTOMATICALLY PROVIDES DROPDOWN ARROWS NEXT TO THE FIELD NAMES THAT ALLOW YOU TO FILTER.
- THERE ARE ALSO THREE WAYS TO FILTER ON ANY RECTANGULAR DATA SET WITH VARIABLE NAMES:
    1. USE THE FILTER BUTTON FROM THE SORT & FILTER DROPDOWN LIST ON THE HOME RIBBON.
    2. USE THE FILTER BUTTON FROM THE SORT & FILTER GROUP ON THE DATA RIBBON.
    3. RIGHT-CLICK ANY CELL IN THE DATA SET AND SELECT FILTER. YOU GET SEVERAL OPTIONS, THE MOST POPULAR OF WHICH IS FILTER BY SELECTED CELL'S VALUE.

CATALOG MARKETING.XLSX

- **OBJECTIVE**: TO INVESTIGATE THE TYPES OF FILTERS THAT CAN BE APPLIED TO THE HYTEX DATA.
- **SOLUTION**: THERE IS ALMOST NO LIMIT TO THE FILTERS YOU CAN APPLY, BUT HERE ARE A FEW POSSIBILITIES:
  - FILTER ON ONE OR MORE VALUES IN A FIELD.
  - FILTER ON MORE THAN ONE FIELD.
  - FILTER ON A CONTINUOUS NUMERICAL FIELD.
  - *TOP 10* AND *ABOVE/BELOW AVERAGE* FILTERS.
  - FILTER ON A TEXT FIELD.
  - FILTER ON A DATE FIELD.
  - FILTER ON COLOR OR ICON.
  - USE A CUSTOM FILTER.

EXAMPLE 2.7
CATALOG MARKETING.XLSX

RESULTS FROM A TYPICAL FILTER

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Person	Age	Gender	Own Home	Married	Close	Salary	Children	History	Catalogs	Region	State	City	First Purchase	Amount Spent
155	154	2	0	1	1	0	$96,800	3	NA	24	Midwest	Kentucky	Louisville	4/28/2012	$3,082
163	162	2	0	1	1	1	$62,200	3	NA	24	Midwest	Indiana	Indianapolis	6/7/2008	$2,119
245	244	2	1	1	1	0	$82,400	2	3	24	Midwest	Indiana	Indianapolis	3/25/2011	$2,035
370	369	2	1	1	1	0	$113,400	3	3	18	Midwest	Kentucky	Louisville	11/25/2011	$1,790
430	429	2	1	1	1	1	$113,000	2	2	18	Midwest	Kentucky	Louisville	6/15/2011	$1,554
570	569	2	1	1	1	1	$70,400	2	NA	12	Midwest	Indiana	Indianapolis	4/12/2007	$1,127
764	763	2	0	1	1	1	$85,500	2	2	18	Midwest	Kentucky	Louisville	7/3/2011	$895
790	789	2	1	1	1	1	$74,500	2	2	12	Midwest	Indiana	Indianapolis	3/7/2012	$824
804	803	2	0	1	1	1	$72,200	2	2	18	Midwest	Kentucky	Louisville	5/29/2011	$715
851	850	2	1	1	1	1	$77,100	2	2	6	Midwest	Indiana	Indianapolis	6/17/2012	$568
1002	Total						$84,750								$14,709