BUSINESS ANALYTICS II

EXPLORATORY DATA ANALYSIS

- The process of statistics starts when we identify what group we want to study or learn something about.
- This group is called the population where it refers to the entire group you want to focus upon.
- In most cases a population is too large to study.
- A more practical approach is to examine the data only from a sub-group of the population, which we can call a sample.
- The sample should represent the population.

EXPLORATORY DATA ANALYSIS

- Exploratory data analysis often reveals new ways to think about the data.
- Exploratory data analysis helps scientists refine their questions and sometimes even reveal entirely new questions.

DATA SET

- Observations organized into variables.
- Set of data made up of individual observations and variables.
- Typically displayed in tables in which rows are observations and columns are variables.
- The data set we will use in this module is the U.S. National Epidemiological Survey on Alcohol and Related Conditions (NESARC). It is a survey designed to determine the magnitude of alcohol use and psychiatric disorders in the U.S. population. It is a representative sample of the non-institutionalized population 18 years and older.

CODEBOOK

- A codebook provides details on each variable within a dataset.
- Explains the data type, size, optionality, range of values etc. for each variable in the data set.
- You will be provided with codebooks to choose from for the C.A worth 70%.
- Sometimes called data dictionaries, codebooks offer complete information on the data set.
- Reviewing a codebook is always the first step in research based on existing data.

CODEBOOK

- Codebooks are used first to help generate a research question.
- Data is often useless and uninterpretable without them.
- Response options are listed and sometimes the frequencies of each response.
- In the following example the variable responses are numeric so rather than showing the responses and their frequencies the code book includes a description of how the variable is measured. (degrees, km etc.)

CODEBOOK

Variables with Descriptions

- CRATER_ID - crater ID for internal sue, based upon the region of the planet (1/16ths), the "pass" under which the crate was identified, ad the order in which it was identified
- LATITUDE_CIRCLE_IMAGE - latitude from the derived center of a non-linear least-squares circle fit to the vertices selected to manually identify the crater rim (units are decimal degrees North)
- LONGITUDE_CIRCLE_IMAGE - longitude from the derived center of a non-linear least-squares circle fit to the vertices selected to manually identify the crater rim (units are decimal degrees East)
- DIAM_CIRCLE_IMAGE - diameter from a non-linear least squares circle fit to the vertices selected to manually identify the crater rim (units are km)
- DEPTH_RIMFLOOR_TOPOG – average elevation of each of the manually determined N points along (or inside) the crater rim (units are km)

CODEBOOK

Gapminder Codebook

Variable Name	Description of Indicator	Main Source
incomeperperson	2010 Gross Domestic Product per capita in constant 2000 US$. The inflation but not the differences in the cost of living between countries has been taken into account.	World Bank Work Development Indicators
alcconsumption	2008 alcohol consumption per adult (age 15+), litres Recorded and estimated average alcohol consumption, adult (15+) per capita consumption in litres pure alcohol	WHO
armedforcesrate	Armed forces personnel (% of total labor force)	Work Development Indicators
breastcancerper100TH	2002 breast cancer new cases per 100,000 female Number of new cases of breast cancer in 100,000 female residents during the certain year.	ARC (International Agency for Research on Cancer)
co2emissions	2006 cumulative CO2 emission (metric tons), Total amount of CO2 emission in metric tons since 1751.	CDIAC (Carbon Dioxide Information Analysis Center)
femaleemployrate	2007 female employees age 15+ (% of population) Percentage of female population, age above 15, that has been employed during the given year.	International Labour Organization
employrate	2007 total employees age 15+ (% of population) Percentage of total population, age above 15, that has been employed during the given year.	International Labour Organization
HIVrate	2009 estimated HIV Prevalence % - (Ages 15-49) Estimated number of people living with HIV per 100 population of age group 15-49.	UNAIDS online database
Internetuserate	2010 Internet users (per 100 people) Internet users are people with access to the worldwide network.	World Bank
lifeexpectancy	2011 life expectancy at birth (years) The average number of years a newborn child would live if current mortality patterns were to stay the same	1. Human Mortality Database, 2. World Population Prospects: 3. Publications and files by

NESARC CODEBOOK

Tape Location	Source Code	Frequency	Item value and description
115-115	FMARITAL		IMPUTATION FLAG FOR MARITAL STATUS
		43038	0. No
		55	1. Yes
116-117	S1Q3B		NUMBER OF MARRIAGES (NOT COUNTING LIVING TOGETHER AS IF MARRIED)
		42941	0-14. Number of Marriages
		152	99. Unknown
118-119	S1Q4A		AGE AT FIRST MARRIAGE
		31794	14-94. Age
		543	99. Unknown
		10756	BL. NA, never married
120-120	S1Q4B		HOW FIRST MARRIAGE ENDED
		4025	1. Widowed
		10803	2. Divorced
		201	3. Other
		98	9. Unknown
		27966	BL. NA, never married; still married to 1st spouse (including legally separated)
121-122	S1Q4C		AGE WHEN STOPPED LIVING WITH FIRST SPOUSE/FIRST SPOUSE DIED
		15498	14-96. Age
		733	99. Unknown
		26862	BL. NA, never married; still married to and living with 1st spouse
123-124	S1Q4D		AGE WHEN MARRIED CURRENT SPOUSE
		20493	14-85. Age
		276	99. Unknown
		22324	BL. NA, not currently married
125-126	S1Q5A		NUMBER OF CHILDREN EVER HAD, INCLUDING ADOPTIVE, STEP AND FOSTER CHILDREN
		42632	0-14. Number of Children
		53	15. 15 or more
		408	99. Unknown

RESEARCH QUESTION

- Once you select a data set of interest to you then you can identify a topic of interest and print those pages of the codebook that include the variable or variables that measure your selected topic.
- Many codebooks are too large to print so it can be important to create your own code book with only those pages that include the variables that you want to examine.

EXPLORATORY DATA ANALYSIS

- Converting raw data into contextual data or data that "makes sense"
  - Organising and summarising raw data
  - Looking for important features and patterns
  - Looking for deviations in the patterns
  - Interpreting your findings

ORGANISE AND SUMMARISE

- Summarise and investigate the distribution of variables of interest to us.
- Investigate one variable at a time: univariate analysis.
- Examine and summarise the distribution of each variable, what values occur and how often?
- Answer research question, tell a story using the variables of interest.

EXAMPLE

- 1200 college students
  - What is your perception of your own body?
    - Underweight, overweight, about right
  - How many students fall into each category?
  - Are they equally distributed?
  - Summarise how often each category occurs.
  - Category value, count and percentage (how often)

SUMMARISE

- It's difficult answer these questions by looking at the raw data, that is the 1200 responses.
- These questions will be easily answered once we summarise and look at the frequency distribution of the variable body image.
- We summarise how often each of the categories occurs.

SUMMARISING

- Summarising as below:

Category	Count	Percent
About right	855	$(\frac{855}{1200}) x 100 = 71.3\%$
Overweight	235	$(\frac{235}{1200}) x 100 = 19.6\%$
Underweight	110	$(\frac{110}{1200}) x 100 = 9.2\%$
Total	n=1200	100.00%

- Next we interpret in the context of the research question

INTERPRETATION

- What % of the sample students fall into each category?
- How are students divided across the categories?
- Are they equally divided?

EXAMPLE

- Research question:
- Are smoking and nicotine dependence associated?
- This question has not changed just refined to:
- Are smoking and nicotine dependence associated among young adults who have smoked in the past 12 months?

NESARC DATA SET

- Nesarc data set and code book are used throughout the labs for this module.
- The topic of interest for the labs is nicotine dependence.
- Interest in exploring the association between smoking behavior and nicotine dependence.
- Identify variables that measure smoking behavior such as smoking status, usual frequency and usual quantity.