GENERAL ASSEMBLY

DATA MANIPULATION

Angelo Klin

Katra Analytics

LEARNING OBJECTIVES

- Review about cleaning the Data

- Review the value of Exploratory Data Analysis

OPENING

CLEANING DATA

DATA FORMAT

DATA FORMAT

- As a result of experiments or observations data is **commonly** presented as:
  - Rows, one for each case, observation, subject
  - Columns
    - Identifiers, subject id; date, time or timestamp
    - Explanatory variables
    - Outcome variable (sometimes absent)
- Stored on paper, spread sheets, databases or text files

MISSING VALUES

MISSING VALUES

- MCAR, Missing Completely at Random
  - does not depend on the variable of interest or any other variable in the dataset
  - very rarely found and the best method is to ignore such cases

- MAR, Missing at Random
  - X_i is missing at random if missingness does not depend on the value of X_i after controlling for another variable

MISSING VALUES

MISSING VALUES

- NAMR, Not missing at Random
  - the missingness mechanism depends on the actual value of missing data
  - modelling such a condition is a very difficult task to achieve
  - the only way to attain an estimate of parameters is to model the missingness, meaning to write a model for missing data and then integrate it back
  - easier said than done

OPENING

EXPLORATORY DATA ANALYSIS

ROOT CAUSE FOR EDA

ROOT CAUSE FOR EDA

- Column of numbers are difficult to read, especially in large volumes, and so determining important characteristics of the data

- Exploratory Data Analysis techniques have been devised as an aid

- The techniques work in part by hiding certain aspects of the data while making other aspects more clear

MAIN REASONS FOR EDA

MAIN REASONS FOR EDA

- Detection of mistakes
- Checking of assumptions
- Preliminary selection of appropriate models
- Determining relationships among the explanatory variables
- Assessing the direction and rough size of relationships between explanatory and outcome variables

- Most of data handling that is not formal statistical modelling and inference can be considered exploratory data analysis

EDA CLASSIFICATION

EDA CLASSIFICATION

- Presentation
  - Non-Graphical, computation of statistics
  - Graphical, uses charts, diagrams and visual resources

- Scope
  - Univariate, each variable by itself, one at a time
  - Multivariate (usually Bivariate), look for relationships amongst the variables

- There are further divisions based on the variable’s role (outcome or explanatory) and type (categorical or quantitative)

UNIVARIATE NON-GRAPHICAL EDA

UNIVARIATE NON-GRAPHICAL EDA

- The observations or measurements make a sample distribution

- Useful to understand the population

- The usual goals of univariate non-graphical EDA
  - to better appreciate the “sample distribution”
  - to make some tentative conclusions about what population distribution(s) is/are compatible with the sample distribution
  - to detect outliers

UNIVARIATE NON-GRAPHICAL EDA - CATEGORICAL DATA

UNIVARIATE NON-GRAPHICAL EDA - CATEGORICAL DATA

- The characteristics of interest for a categorical variable are
  - the range of values
  - the frequency (or relative frequency) of occurrence for each value
- For ordinal variables it is sometimes appropriate to treat them as quantitative variables
- The only useful techniques is some form of tabulation of the frequencies, usually along with calculation of the fraction (or percent) of data that falls in each category

UNIVARIATE NON-GRAPHICAL EDA - CATEGORICAL DATA

UNIVARIATE NON-GRAPHICAL EDA – CATEGORICAL DATA

- Losing data is a common mistake and EDA is very helpful for finding mistakes

- Expect that the proportions add up to 1.00 (or 100%) if the calculations are correct (count/total)

	A	B	C	Other	Total
Count	5	6	4	5	20
Proportion	0.25	0.30	0.20	0.25	1.00
Percent	25%	30%	20%	25%	100%

UNIVARIATE NON-GRAPHICAL EDA - QUANTITATIVE DATA

UNIVARIATE NON-GRAPHICAL EDA – QUANTITATIVE DATA

- Characteristics of a quantitative variable of a population distribution
  - centre
  - spread
  - modality (number of peaks in the PDF (Probability Density Function))
  - shape (including "heaviness of the tails")
  - outliers
- Observed data represent just one sample out of an infinite number of possible samples
  - The characteristics of a randomly observed sample are not inherently interesting, except to the degree that they represent the population that it came from

BIVARIATE NON-GRAPHICAL EDA

BIVARIATE NON-GRAPHICAL EDA

- Categorical data (and quantitative data with only a few different values)
  - Cross-tabulation for two variables
    - a two-way table with column headings that match the levels of one variable
    - row headings that match the levels of the other variable
    - the counts of all subjects that share a pair of levels
  - The two variables might be both explanatory, both outcome or one of each
  - Row percentages (which add to 100% for each row), column percentages (which add to 100% for each column) and cell percentages (which add to 100% over all cells) are also useful

BIVARIATE NON-GRAPHICAL EDA

BIVARIATE NON-GRAPHICAL EDA

- Cross-tabulation

	Female	Male	Total
Young	2	3	5
Middle	2	1	3
Old	3	0	3
Total	**7**	**4**	**11**

HOW TO MAKE A BAD GRAPH

HOW TO MAKE A BAD GRAPH

- The aim of good data graphics
    - Display data accurately and clearly

- Some rules for displaying data badly
    - Display as little information as possible
    - Obscure what you do show (with chart junk)
        - Use pseudo-3D and colour gratuitously
    - Make a pie chart (preferably in colour and 3D)
        - Use a poorly chosen scale

EDA RECAP

EDA RECAP

- In a nutshell
  - You should always perform appropriate EDA before further analysis of your data
  - Perform whatever steps are necessary to become more familiar with your data
  - check for obvious mistakes
  - learn about variable distributions and learn about relationships between variables

- EDA is not an exact science - it is a very important art!

CONCLUSION

TOPIC REVIEW

TOPIC REVIEW

- Data Manipulation to fix the data
- Exploratory Data Analysis to understand the data

Q & A