MDB Notes

Matthew Beckman

6/28/2021

MB Intro

- Why interested in the topic
  - IMO suffers a wide gap between real & percieved value for many students
  - careful EDA democratizes data analysis IMO (curiosity & persistence are the main prereqs)
  - EDA seems ripe for opportunity
- Stubborn pattern that EDA is a cursory obligation to print 5 number summaries & a few boxplots
- Courses & student populations
  - typically statistics majors are front of mind
  - Senior capstone course for statistics majors

STAT 184 EDA Guidance

1. **Examine the data source:**
   - variable types,
   - coding,
   - missingness,
   - summary statistics/plots,
   - who/what/when/where/why/how data were collected (e.g., motivations, circumstances, & provenance for data)
2. **Discover features that influence may modeling decisions:**
   - investigate potential outliers,
   - consideration for recoding variables (e.g., numeric data that's functionally dichotomous),
   - evaluate correlation structure (e.g., autocorrelation, hierarchy, spatial/temporal proximity)
3. **Address research questions:**
   - build intuition and note preliminary observations/conclusions related to RQ's as stated
   - note observations that prompt you to refine your research questions or add new questions to investigate

Breakout 2 (common EDA resources)

Toward Data Science

Shan, T. (2020). An extensive step by step guide to exploratory data analysis. *Towards data science.* URL: https://towardsdatascience.com/an-extensive-guide-to-exploratory-data-analysis-ddd99a03199e

Key Quotes

"[EDA] is a step in the Data Analysis Process, where a number of techniques are used to better understand the dataset" e.g.,

- Discern important variables from useless variables
- Search for outliers, missing values, or human error
- Study relationships between variables (or lack thereof)
- maximize insights from data and mitigate issues/errors that may occur in subsequent analysis

"Many EDA techniques can remedy some common problems that are present in every dataset"

"Exploratory Data Analysis does two main things: 1. It helps clean up a dataset. 2 It gives you a better understanding of the variables and the relationships between them."

"[T]here are [three] main components of exploring data: 1. Understanding your variables 2. Cleaning your dataset 3. Analyzing relationships between variables."

Worked Example (Used Car Data)

1. Understanding your variables:
    - examine dimensions: 525,839 rows, 22 columns
    - print variable names
    - examine summary statistics for each variable
        - identify apparent errors in the data ($3 billion used car? 10 million miles on odometer)
        - notes a variable with high rate of missingness (condition of the car)
2. Cleaning your dataset
    - remove redundant variables
    - remove variables with 40% or more missingness
    - remove outliers according to intuition about price, year, odometer, etc
    - remove rows with null values
    - result: 208,765 rows of data
3. Analyzing relationships
    - correlation matrix (all pairs)
    - scatterplot matrix (all pairs)
    - histogram and/or boxplot of each variable

MDB Remarks

- one-size fits all approach... "personal guide to performing EDA for any dataset"
- no apparent RQ in the beginning
- no comment on where the data have come from and/or whether these data are suitable for any particular use.

- no comment on 60% data loss during cleaning
- summaries shown aren't especially informative
- there had been an earlier nod to curiosity in an anecdote about football data (Google search TE vs LB), but nothing of the sort in the worked example

NIST–Engineering Statistics Handbook

NIST/SEMATECH (2021). *e-Handbook of Statistical Methods.* Exploratory Data Analysis. URL: https://www.itl.nist.gov/div898/handbook/eda/eda.htm Accessed: 28 June 2021.

Key Quotes

"EDA is an approach to data analysis that postpones the usual assumptions about what kind of model the data follow with the more direct approach of allowing the data itself to reveal its underlying structure and model. EDA is not a mere collection of techniques; EDA is a philosophy as to how we dissect a data set; what we look for; how we look; and how we interpret."

"The reason for the heavy reliance on graphics is that by its very nature the main role of EDA is to open-mindedly explore, and graphics gives the analysts unparalleled power to do so, enticing the data to reveal its structural secrets, and being always ready to gain some new, often unsuspected, insight into the data."

"How does [EDA] differ from summary analysis? EDA has as its broadest goal the desire to gain insight into the engineering/scientific process behind the data. Whereas summary statistics are passive and historical, EDA is active and futuristic. In an attempt to 'understand' the process and improve it in the future, EDA uses the data as a 'window' to peer into the heart of the process that generated the data."

"EDA is a data analysis approach... Three popular data analysis approaches are: 1 Classical; 2 Exploratory (EDA); 3 Bayesian. Data analysts freely mix elements of all of [all] three approaches (and other approaches)."

"The primary goal of EDA is to maximize the analyst's insight into a data set and into the underlying structure of a data set, while providing all of the specific items that an analyst would want to extract from a data set, such as:

1. a good-fitting, parsimonious model
2. a list of outliers
3. a sense of robustness of conclusions
4. estimates for parameters
5. uncertainties for those estimates
6. a ranked list of important factors
7. conclusions as to whether individual factors are statistically significant

8. optimal settings"

“it is not enough for the analyst to know what is in the data; the analyst also must know what is not in the data, and the only way to do that is to draw on our own human pattern-recognition and comparative abilities in the context of a series of judicious graphical techniques applied to the data.”

MDB Remarks

- emphasis on graphical analysis
- emphasis on curiosity and open-minded exploration
- emphasis on scientific question
- framed as a method of data analysis, and even contrasted with Classical & Bayesian analysis.