Proposal Abstract

Matthew Beckman

1/29/2021

Theme: Expanding opportunities https://www.causeweb.org/cause/uscots/uscots21/about

Title

- Is Exploratory Data Analysis Dying When We Need it Most?

- Time to upgrade EDA
- Rethinking EDA in Data Science

Abstract (150 words)

Exploratory Data Analysis (EDA) perhaps suffers from a wider gap between real & perceived value than any contribution the data analyst offers. A careful EDA–i.e., initial data analysis, in particular–requires critical thinking and creativity, while building fruitful intuition for tentative research questions and laying foundations for responsible exploration, inference, and prediction. However, EDA too often seems confined to a cursory obligation between data preparation (the 'boring' part) and modeling/machine learning (the 'exciting' part).

The popularity of Data Science has ushered in open data initiatives and elegant software solutions yield easy access to models and algorithms that can be readily implemented by a non-specialist to explore all manner of subjects, even if the workings, assumptions, and proper interpretations of those methods are quite opaque to the user. This session aims to provoke a discussion that seeks to (1) reframe/affirm goals of a thorough EDA in a world laden with found data, big data, etc. (2) rethink where/how EDA should be addressed in data analysis courses at all levels.

148 words

Exploratory Data Analysis (EDA) perhaps suffers a wider gap between real & perceived value than any contribution the data analyst offers. The popularity of Data Science has ushered in a rising tide of open-data initiatives and elegant software solutions readily accessible to non-specialists, thereby democratizing data analysis for all manner of subjects. Initially, some novices may feel overwhelmed and others may risk spurious results utilizing opaque methods with fancy names, yet all can be empowered by careful EDA based on lucid methods powered by curiosity, creativity, and critical thinking to build fruitful intuition and responsible insights. However, EDA too often seems a cursory obligation to be minimized.

This session aims to provoke a discussion that seeks to (1) recast/affirm goals of a thorough EDA in a world laden with found data, big data, etc. (2) rethink where/how EDA should be addressed in data analysis courses at all levels.

Goals / take-aways

- What should a thorough EDA include?
- Data science has helped to shift attention toward a wider view of data (volume, variety, etc). Which (if any) priorities for EDA need to adapt, and which are unchanged?
- Do we need to rethink where and how EDA should be addressed in Statistics & Data Science courses

Connection to Theme

In a very real sense, I would argue that EDA democratizes data analysis. The prerequisites creativity and critical thinking rather than fancy models with sophisticated underpinnings. Through careful attention to EDA, students can be equipped with powerful tools for learning from data from the first weeks of a first course with very few technical obstacles. Open-data initiatives (e.g., data.gov) and remote hosting services (e.g., GitHub, Kaggle) have made all manner of data on an enormous range of subjects available to the public, and increasingly powerful software tools are designed to enable the non-specialist (as well as the specialist) to explore data to build intuition for sophisticated questions. In many cases, a thoughtful framework for responsible EDA may well differentiate whether the combination of rich open data and novice-friendly tools results in fruitful citizen (data) science or spurious conclusions that loosely appeal to the authority of "the data."

Engagement (Agenda & interactive elements)

- (10 min) Opening remarks to share some existing perspectives on the role and contributions of EDA and IDA–i.e., initial data analysis
- (10 min) small group discussion prompt(s)–e.g.,
  - Remarks on assertion of gap between real & perceived value of EDA?
  - What should a thorough EDA include?
- (10 min) large group discussion based on observations from small groups
- (10 min) small groups review & discuss various guidance for EDA–e.g.,
  - Velleman & Hoaglin (2012). Exploratory data analysis. In *APA handbook of research methods in psychology* <doi.org/10.1037/13621-003>
  - Shan (2020). An extensive step by step guide to exploratory data analysis. *Towards data science.*
  - Wickham & Grolemund (2020). *R for Data Science* <r4ds.had.co.nz>
- (10 min) large group discussion to compare & contrast observations from small group discussions. Especially with respect to revisions or new insights about what a thorough EDA should include?
- (10 min) small group discussions prompt(s)–e.g.,
  - Data science has helped to shift attention toward a wider view of data (volume, variety, etc). Which (if any) priorities for EDA need to adapt, and which are unchanged?
  - Do we need to rethink where and how EDA should be addressed in Statistics & Data Science courses
- (15 min) large group discussion based on observations from small groups