Exploratory Data Analysis with Grammar of Graphics

School of Information Sciences
Manipal University

Exploratory Data Visualization

What are the elementary perceptual tasks that people perform in obtaining quantitative information from data graphics?

> accuracy of the extracted quantitative information is NOT the ONLY important aspect of data graphics.

Exploratory Data Analysis

An approach or philosophy for data analysis

- employs a variety of techniques, mostly graphical or visual

Exploratory Data Analysis - Origins

John Tukey of AT&T Bell Labs was the original champion of exploratory data analysis (EDA).

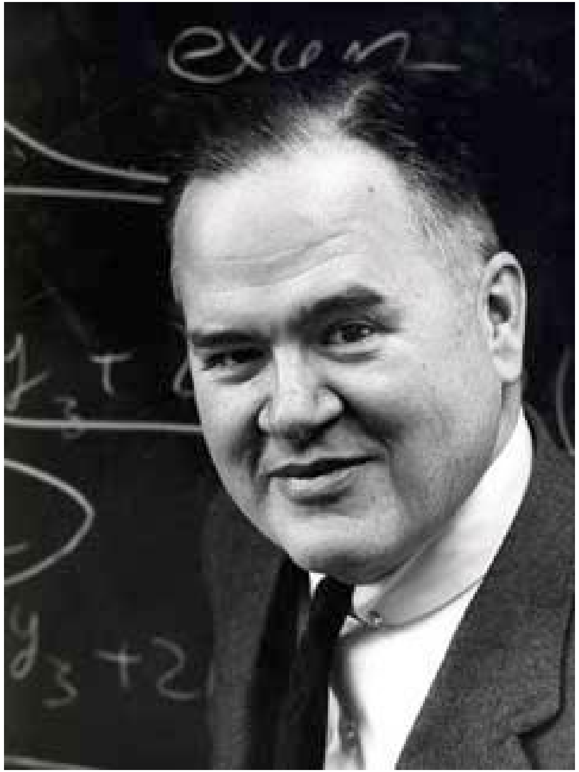
- encourage statisticians to explore the data.
- possibly formulate hypotheses leading to new data collection and experiments.

Exploratory Data Analysis - Writings

John Tukey

1961 - The Future of Data Analysis

1977 - Exploratory Data Analysis
Addison-Wesley Publishing



Exploratory Data Analysis - Tools

John Tukey influenced the development of an early statistical package called S

- family of statistical-computing environments featured vastly improved dynamic visualization capabilities.
- allowed statisticians to identify outliers, trends and patterns in data that merited further study.
- Programming with Data

History of Exploratory Data Analysis

Spring of 1976

The earliest beginnings of S came from discussions in the among a group of five people at Bell Labs.

Rick Becker
John Chambers
Doug Dunn
John Tukey and
Graham Wilkinson

Evolution of S

http://ect.bell-labs.com/sl/S/history.html

"We were looking for a system to support the research and the substantial data analysis projects in the statistics research group at Bell Labs. This motivation was different from either the perspective of a service organization or of an individual researcher in an academic situation, although we shared some of the concerns of each of those situations."

John Chambers

Evolution of S

http://ect.bell-labs.com/sl/S/history.html

“We were concerned to support serious data analysis (although for some time some of our colleagues were skeptical about serious analysis from an interactive system). However, little or none of our analysis was standard, so flexibility and the ability to program were essential from the start.”

Rick Becker

ACM Software System for S

ACM honors Dr. John M. Chambers of Bell Labs with the 1998 ACM Software System award for creating “S System” software.

http://oldwww.acm.org/announcements/ss99.html

John Chambers donated the prize money (US$10,000) to the American Statistical Association to endow an award for novel statistical software.

http://stat-computing.org/awards/jmc/history.html

Exploratory Data Analysis - Objectives

- Maximize insight into a dataset
- Extract important variables
- Uncover underlying structure
- Detect outliers and anomalies
- Test hypothesis and assumptions
- Develop parsimonious models
- Determine optimal factor settings

Exploratory Data Analysis - Objectives

- Suggest hypotheses about the causes of observed phenomena.
- Assess assumptions on which statistical inference will be based.
- Support the selection of appropriate statistical tools and techniques.
- Provide a basis for further data collection through surveys or experiments.

Paradigms of Analysis Techniques

Classical analysis

Problem => Data => Model => Analysis => Conclusions

Exploratory Data Analysis

Problem => Data => Analysis => Model => Conclusions

Bayesian method

Problem => Data => Model => Prior Distribution => Analysis => Conclusions

Data Visualization

Grammar of Graphics

Grammar *and* Graphics

Grammar of a language consists of a set of rules defining well-formed sentences of that language.

Graphics deals with pictorial rendering of visually meaningful artifacts.

Grammar *and* Graphics

Grammar of a language consists of a set of rules defining well-formed sentences of that language.

- Humans learn grammar.

Graphics deals with pictorial rendering of visually meaningful artifacts.

- Human visual perception system is a complex system.
- It is amenable to training and conditioning.

Grammar *and* Graphics

Grammar of a language consists of a set of rules defining well-formed sentences of that language.

Graphics deals with pictorial rendering of visually meaningful artifacts.

- Computer graphics defines computational methods for rendering semantically rich and interactive graphics.
- Takes into account capabilities and limitations of human visual perception.

The Role of Graphics in Data Analysis

Statistics and data analysis procedures can ve broadly classified as

- quantitative
- graphical

The Role of Graphics in Data Analysis

Quantitative techniques are statistical procedures that yield numeric or tabular output

- hypothesis testing
- analysis of variance
- point estimates and confidence intervals
- least squares regression

The Role of Graphics in Data Analysis

However, there is a large collection of statistical charts that are generally referred to as graphical methods.

- scatter plots
- histograms
- probability plots
- residual plots
- box plots
- block plots, and *many more…*

The Role of Graphics in Data Analysis

EDA exploits graphical tools to gain insight into a data set

- testing assumptions
- model selection
- model validation
- estimator selection
- relationship identification
- factor effect determination
- outlier detection

Grammar *of* Graphics

A set of rules to express rich data visualization using computers as powerful *medium*.

> The *message* is visually communicated.
>
> The message is *interactively* communicated.

Tenets of Grammar of Graphics

Grammar of Graphics shuns chart typologies.

GoG considers charts as instances of much more general objects.

- a vocabulary based solely on charts will only offer fewer charts than people really want.
- The reportoire becomes difficult to extend (if not impossible to extend)

Three Stages of Graphics Creation

Grammar of Graphics identifies three interacting stages

- Specification
  - ‘graph’ → mathematical object
- Assembly
  - ‘graphics’ → concrete representation of a graph
- Display
  - ‘Aesthetics’ → rendering aspects of graphics

Graphical Specification

DATA

TRANS

SCALE

COORD

ELEMENT

GUIDE

Graphical Specification

DATA	operations that create variables from datasets
TRANS	variable transformations
SCALE	scaling tranformations
COORD	define a coordinate system
ELEMENT	points and their aesthetic properties
GUIDE	guides (axes and legends)

Grammar of Graphics

“Some of the rules and graphics in this book may seem self-evident, especially to those who have never written a computer program. Programming a computer exposes contradictions in commonsense thinking, however.”

The Grammar of Graphics.
Leland Wilkinson.

https://www.cs.uic.edu/~wilkinson/

Grammar of Graphics

“Programming a computer exposes contradictions in commonsense thinking, however. And programming a computer to draw graphics teaches most surely the ancient lesson that God is in the details.”

The Grammar of Graphics.
Leland Wilkinson.

https://www.cs.uic.edu/~wilkinson/

Grammar of Graphics - Inspirations

R programming language.

ggplot2	plotting system for R.
dplyr	manipulating data.
tidyr	tidying data.

Hadley Wickham
http://hadley.nz/
https://github.com/hadley

Grammar *of* Graphics - Process

Canvas Model (Basic R)

- start with a blank canvas.
- add graphical elements one step at a time.

Grammar of Graphics Model (ggplot2)

- organize data for purposeful rendering.
- plan data transformation in a series of stages.
- plot graphics as a mapping from data to aesthetics.

Grammar *of* Graphics - Process

Aesthetics

- elements that appeal to visual perception.
- color, geometric shapes, textures, fonts, and such.
- interaction.
- ease of exploration - user experience with graphics.