EDA Workshop

Orientation Session

Do The Math

Chicago, IL
Bangalore, India
www.mu-sigma.com

November 2, 2017

Learning Outcomes

- Go beyond numbers, variables and math and be able to quantify and visualize the business and its inner working through data
  - Get a multi-dimensional view of the business
  - Observe known facts through data
  - Validate hypotheses based on business intuition
- Foster extraction mindset by understanding the art and the science of discovery
- Understand the basic hygiene around EDAs
  - Gauging data quality
  - Know when to start and when to end an EDA
  - Guidelines on how to conduct an EDA and what NOT to do in an EDA

Why EDA?

Descriptive	**Inquisitive**
Predictive	**Prescriptive**

"Doing statistics is like solving **crosswords** except that one cannot know for sure whether one has found the solution"

– John Tukey

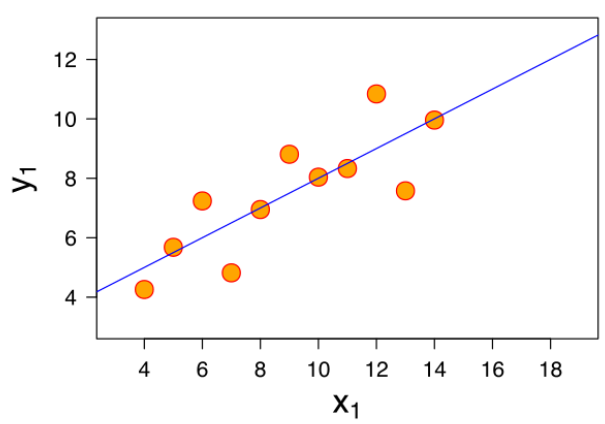
What is EDA? and What is not EDA?

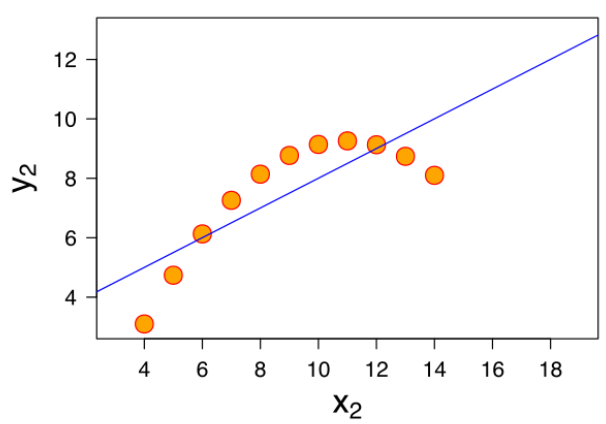
- Getting familiar with the data
  - *Seeing* the data
  - Assess quality of data
  - Hygiene checks, and making the data usable
- Suggesting hypotheses about causes of observed phenomenon
- Assessing assumptions for statistical inference
- Providing a basis for further data collection through surveys and experiments

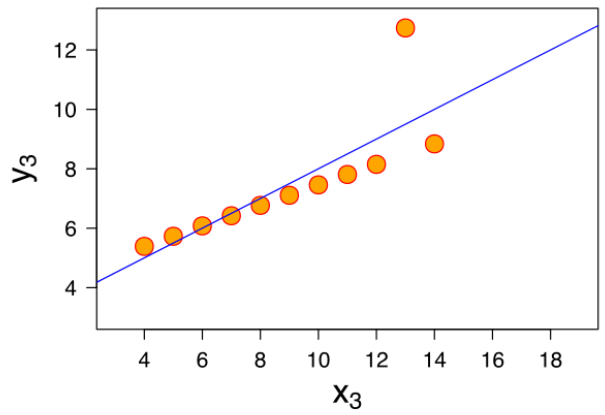
"The **best single device** for suggesting, and at times answering, questions beyond those originally posted is the **graphical display**"

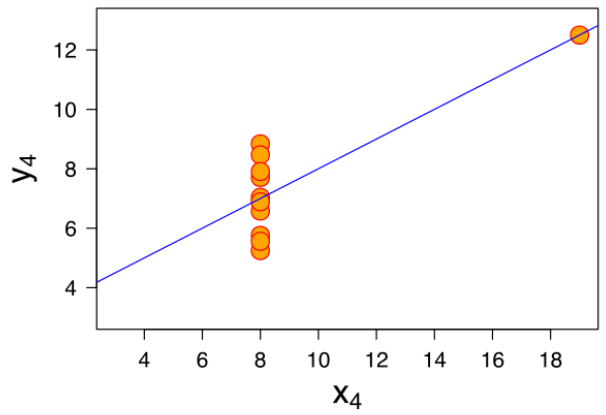
– John Tukey

Anscombe's Quartet – Importance of Visual Representation









Mean of x	9
Sample Variance of x	11
Mean of y	7.5
Sample Variance of y	4.125
Correlation between x and y	0.816

Linear Regression Line
y = 3.00 + 5.00 x

"There is no data that can be displayed in a **pie chart**, that cannot be displayed better in some other type of chart"

– John Tukey

Chart Suggestions—A Thought-Starter

What would you like to show?

Comparison

- Among Items
  - Two Variables per Item
    - Variable Width Column Chart
  - One Variable per Item
    - Many Categories
      - Table or Table with Embedded Charts
    - Few Categories
      - Many Items
        - Bar Chart
      - Few Items
        - Column Chart
- Over Time
  - Many Periods
    - Cyclical Data
      - Circular Area Chart
    - Non-Cyclical Data
      - Line Chart
  - Few Periods
    - Single or Few Categories
      - Column Chart
    - Many Categories
      - Line Chart

Relationship

- Two Variables
  - Scatter Chart
- Three Variables
  - Bubble Chart

Distribution

- Single Variable
  - Few Data Points
    - Column Histogram
  - Many Data Points
    - Line Histogram
- Two Variables
  - Scatter Chart
- Three Variables
  - 3D Area Chart

Composition

- Changing Over Time
  - Few Periods
    - Only Relative Differences Matter
      - Stacked 100% Column Chart
    - Relative and Absolute Differences Matter
      - Stacked Column Chart
  - Many Periods
    - Only Relative Differences Matter
      - Stacked 100% Area Chart
    - Relative and Absolute Differences Matter
      - Stacked Area Chart
- Static
  - Simple Share of Total
    - Pie Chart
  - Accumulation or Subtraction to Total
    - Waterfall Chart
  - Components of Components
    - Stacked 100% Column Chart with Subcomponents

5 sins in analysis

- Theory of relativity - Benchmarking
- Live and Let Live – Unequal Observation Window
- Run vs Drive – Causation vs Correlation
- All that glitters is not gold – Misleading Bivariate
- 2 States – Observation and Performance Window