

mod2

November 10, 2018

0.1 CBA Marathon Courses Fall 2018

1 Data Analysis and Visualization Using Python

1.1 Module 2 - Exploratory Data Analysis (EDA)

1.1.1 Dr. Mohammad AlMarzouq

2 Exploratory Data Analysis

2.1 The act of making sense of data by converting raw data into actionable information

Myatt, Glenn J.; Johnson, Wayne P.. Making Sense of Data I: A Practical Guide to Exploratory Data Analysis and Data Mining. Wiley.

3 Steps In Exploratory Data Analysis

1. Problem definition and planning
2. Data preparation
3. Data analysis
4. Deployment

4 Problem Definition

- Identify the problem to be solved
 - Problem to explore? Question to answer? System to build?
- List project deliverables
 - Report vs System
- Identify required resources/skills and success factors
 - Including data sources
- Assemble team
- Prepare plan

5 Data Preperation (Longest Step)

- Access and combine data
- Summarize data
- Look for errors
- Transform data
- Segment data

6 Data Analysis (Uncertain Step)

- Exploring relationships between variables
- Group summaries and comparisons
- Visualization (**Our focus**)
- Other advanced topics include:
 - Discovering non-trivial patterns
 - Building regression and classification models
 - ... etc

7 Deployment

- Generate report
- Deploy decision-support tool/system
- Measure business impact

8 Notes On The Steps

- They apply to any other advanced type of analysis
- Because the process involves discovery, it is iterative
 - Experience is key
 - Multiple perspective and critical thinking is useful
 - Patience and Perseverance is required

9 Data Visualization

The effort to understand data by placing it in a visual context

10 Dr. Edward Tufte

- Though leader and praticioner of data visualization
- Written two excellent books on the subject:
 - The Visual Display of Quantitative Information
 - Envisioning Information
- Put down some principles for data visualization

11 Excellence in Visualization

- Clear, precise, and efficient communication of complex ideas
- Greatest number of ideas in the smallest amount of time and space
- Multivariate
- Conveys the truth

12 Visualization Goals

- Content focus
- Comparison rather than description
- Integrity
- High resolution
- Utilize designs proven with time

13 The Message

- Can use tables, charts, animations, infographics ..etc
- Powerful if the right data and graphic are [combined](#)
- We will focus mostly on charts and tables, but know that **the possibilities are bigger**.
- To improve your visualization, read the work of **Stephen Few**:
 - Show Me the Numbers: Designing Tables and Graphs to Enlighten
 - Information Dashboard Design: Displaying Data for At-a-Glance Monitoring

14 The Visualization Tools - In Python

- [Matplotlib](#) (We will work with this)
- Bokeh
- ggplot
- [Seaborn](#) (and this)
- Plotly
- Altair
- .. others

They vary in their simplicity and capabilities: static, interactive, animated ..etc.

15 Other Visualization Platforms

- R: ggplot2, ggvis, .. libraries much like python
- [Tableau](#): The current defacto standard in data visualization for non-programmers
- SAS, SPSS, Excel, Matlab, Stata ... etc.

16 The Message: Charts Vs. Tables

- Tables used to accurately show the values of specific data points