EDA - Session 1

Visualization - Exploring the Data I

Content

1. Introduction
2. Visualization
3. Exploring the data I
4. Hands-On:

Introduction

EDA (Exploratory Data Analysis):

"This is the process of navigating through data details and trying to make sense of the individual columns and the relationships between them. "

Introduction

EDA (Exploratory Data Analysis):

"This is the process of navigating through data details and trying to make sense of the individual columns and the relationships between them. "

If you intend to develop machine learning models, having insight into the data can lead to more performant models and understanding why predictions are made.

Introduction

EDA (Exploratory Data Analysis):

In his 1977 book Exploratory Data Analysis, John W. Tukey thinks of EDA as a way to explore data, uncover evidence, and develop hypotheses that can later be confirmed by statistical tests.

Hypothesis-driven vs Data-driven

Visualization

Data visualization is an important aspect of data science, for at least three distinct reasons:

- EDA
- Error detection (insufficient cleaning, outliers, erroneous assumptions)
- Communication

Generally speaking, selected visualization tools depends on the complexity and design, for example: plots for EDA, for publications, for interactive dashboards.

Visualization

Python Visualization Tools:

- **Pyplot Matplotlib**
  - General purpose, procedural and Object Oriented approaches
- **Seaborn**
  - Statistical focus, built on Matplotlib
- **Bokeh, Plotly:**
  - Built on top of JS, dashboards integration.

Exploring the data

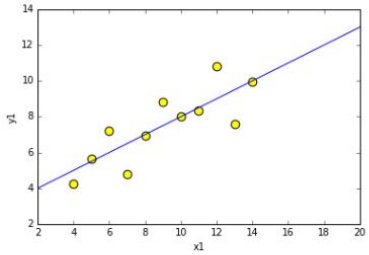
What should you do when encountering a new data set? In general:

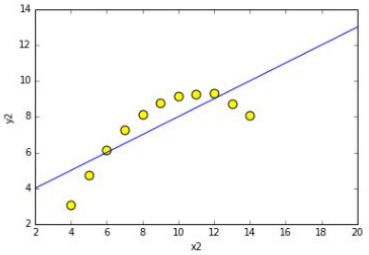
- Answer the basic questions about, context, data set size, fields meaning
- Summary statistics
- Pairwise correlations
- Class breakdowns
- Plots of distributions

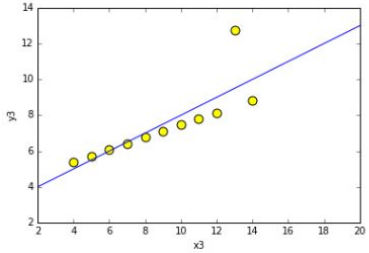
Exploring the data

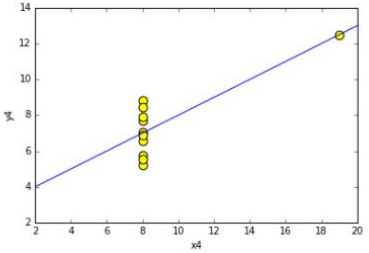
There are limits to how well you can understand data without visualization techniques.

	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
	8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
	13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
	9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
	11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
	14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
	6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
	4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
	12.0	10.84	12.0	9.31	12.0	8.15	8.0	5.56
	7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
	5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89
Mean	9.0	7.5	9.0	7.5	9.0	7.5	9.0	7.5
Var.	10.0	3.75	10.0	3.75	10.0	3.75	10.0	3.75
Corr.	0.816		0.816		0.816		0.816	









These data sets are all dramatically different, even though they have identical summary statistics.

Hands-On