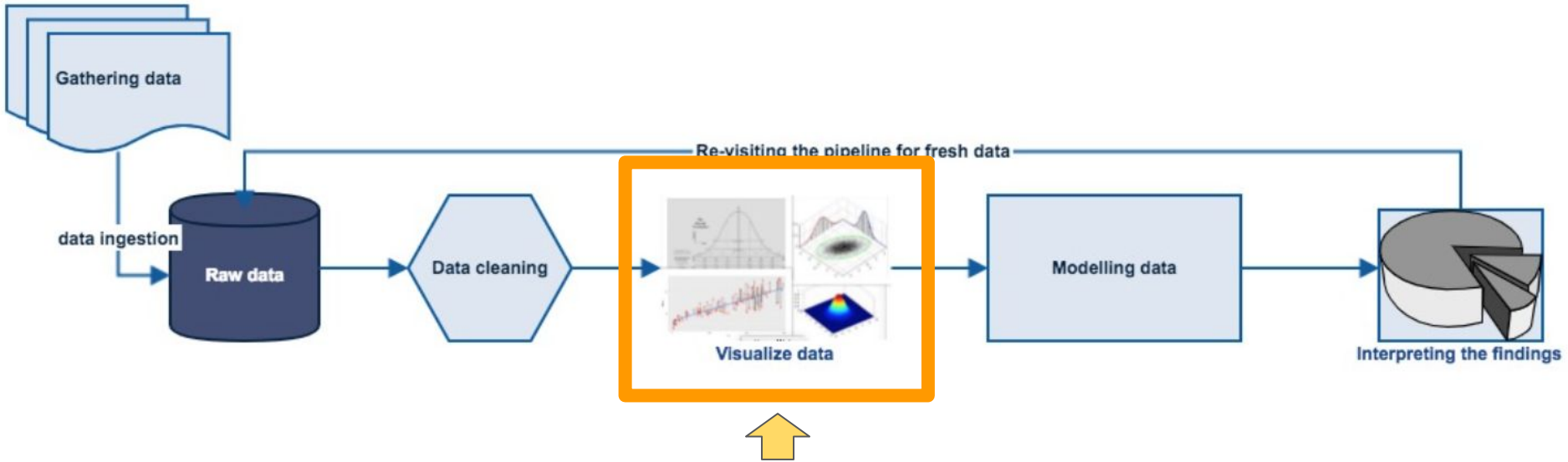
Department of
Computer Science

Exploratory Data Analysis using Seaborn

Erick Gomez Nieto, PhD
emgomez@ucsp.edu.pe

Erick Gomez-Nieto, PhD. Exploratory Data Analysis. Data Science Program.



We are here!

Erick Gomez-Nieto, PhD. Exploratory Data Analysis. Data Science Program.

Exploratory Data Analysis (EDA) is an approach/philosophy for data analysis that employs a variety of techniques (mostly graphical) to

- Maximize insights into a data set.
- Uncover underlying structure.
- Extract important variable
- Detect outliers and anomalies
- Test underlying assumptions
- Develop parsimonious models.
- Determine optimal factor settings.

Source: **NIST/SEMATECH e-Handbook of Statistical Methods**
https://www.itl.nist.gov/div898/handbook/eda/eda.htm

Erick Gomez-Nieto, PhD. Exploratory Data Analysis. Data Science Program.

EDA Goals

- The primary goal of EDA is to maximize the analyst's insight into a data set.
- To get a "feel" for the data, the analyst also must know what is not in the data.
- The only way to do that is to draw on our own human pattern-recognition and comparative abilities in the context of a series of judicious graphical techniques applied to the data.

Erick Gomez-Nieto, PhD. Exploratory Data Analysis. Data Science Program.

EDA philosophy

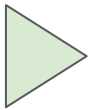
EDA is not identical to statistical graphics although the two terms are used almost interchangeably. Statistical graphics is a collection of techniques--all graphically based and all focusing on one data characterization aspect. EDA encompasses a larger venue;

EDA is an approach to data analysis that postpones the usual assumptions about what kind of model the data follow with the more direct approach of allowing the data itself to reveal its underlying structure and model.

EDA is not a mere collection of techniques; EDA is a philosophy as to how we dissect a data set; **what** we look for; **how** we look; and **how we interpret**. It is true that EDA heavily uses the collection of techniques that we call "statistical graphics", but it is not identical to statistical graphics per se.

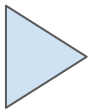
Erick Gomez-Nieto, PhD. Exploratory Data Analysis. Data Science Program.

Paradigms for Analysis Techniques



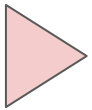
For **classical analysis**, the sequence is

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



For **EDA**, the sequence is

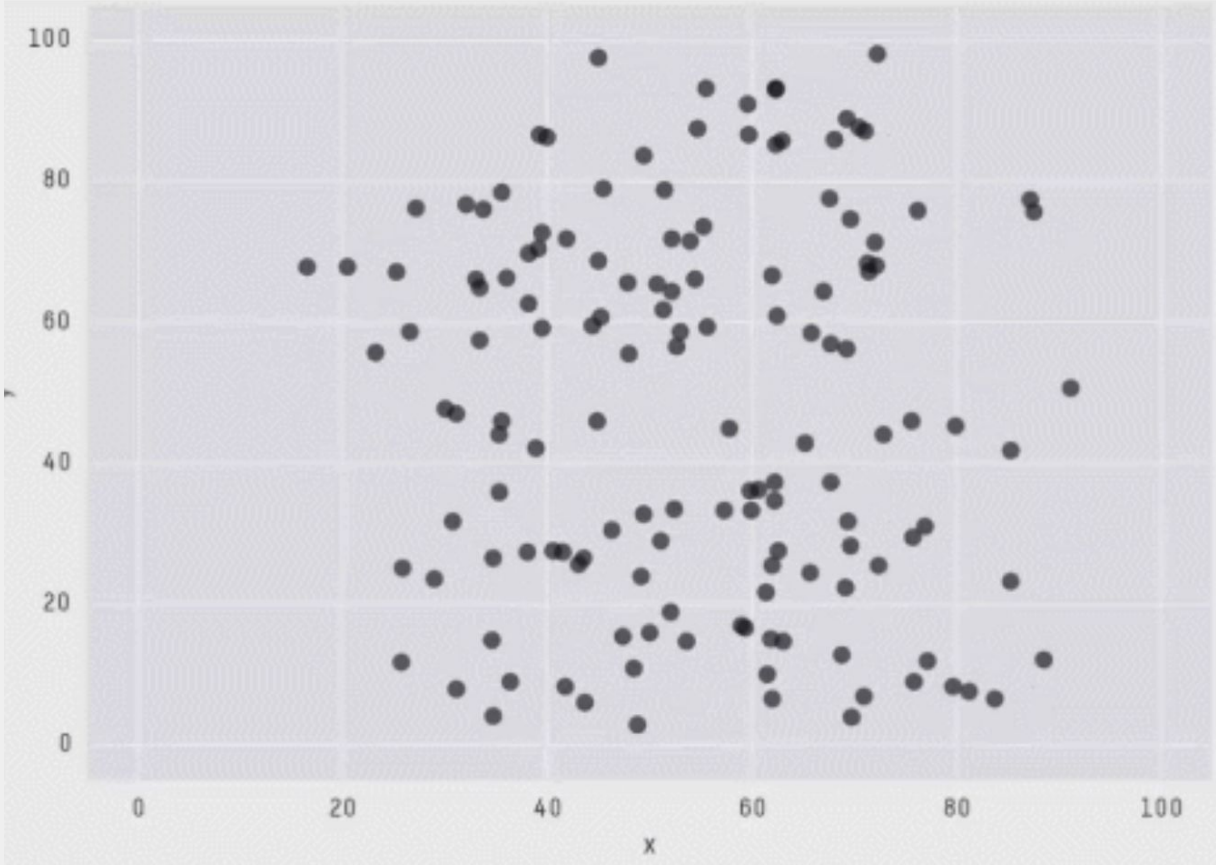
Problem => Data => Analysis => Model => Conclusions



For **Bayesian**, the sequence is

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

Erick Gomez-Nieto, PhD. Exploratory Data Analysis. Data Science Program.

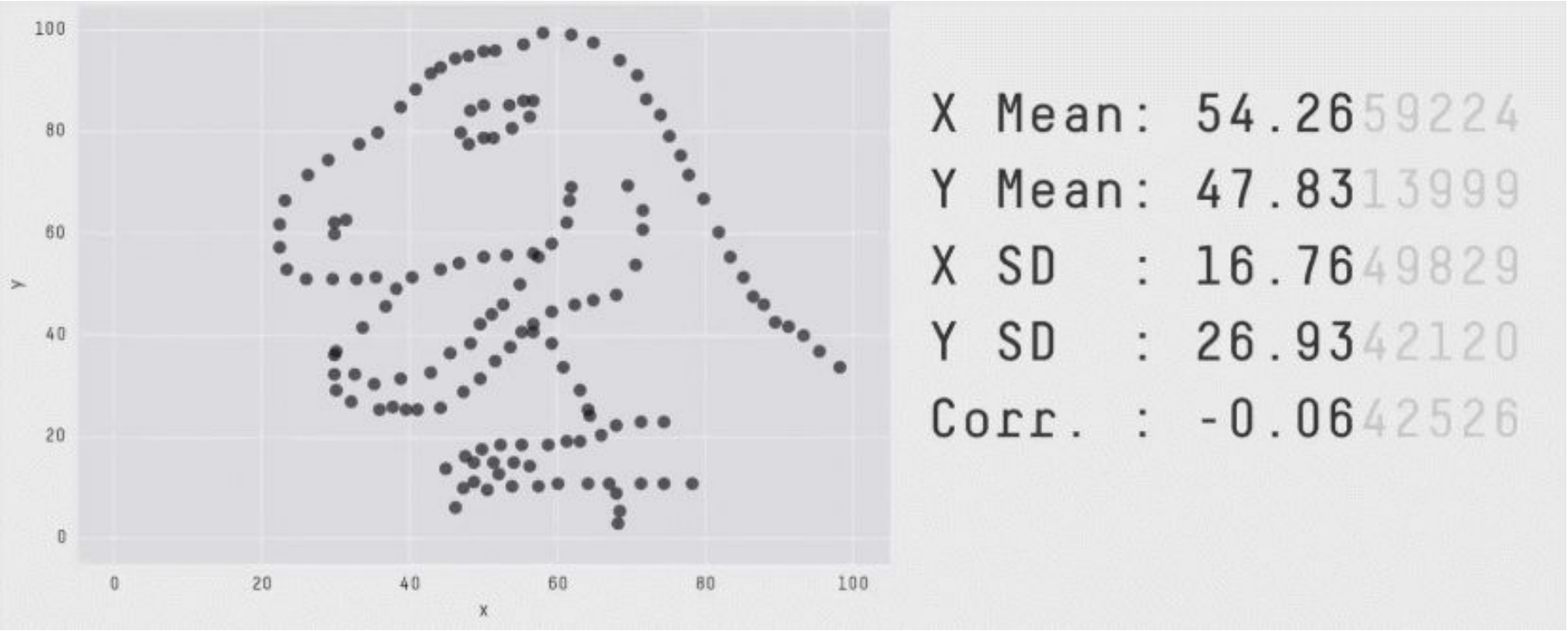


X Mean: 54.26
Y Mean: 47.83
X SD : 16.76
Y SD : 26.93
Corr. : -0.06

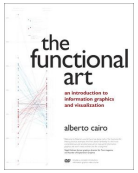
Classical analysis

Erick Gomez-Nieto, PhD. Exploratory Data Analysis. Data Science Program.

The Datasaurus Dozen



Alberto Cairo. 2012. ***The Functional Art: An introduction to information graphics and visualization*** (1st. ed.). New Riders Publishing, USA.



Erick Gomez-Nieto, PhD. Exploratory Data Analysis. Data Science Program.

EDA Approach Relies Heavily on Graphical Techniques

Graphical procedures are not just tools that we could use in an EDA context, they are tools that we must use. Such graphical tools are the shortest path to gaining insight into a data set in terms of

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

If one is not using statistical graphics, then one is forfeiting insight into one or more aspects of the underlying structure of the data.

Source: **NIST/SEMATECH e-Handbook of Statistical Methods**
https://www.itl.nist.gov/div898/handbook/eda/eda.htm

Erick Gomez-Nieto, PhD. Exploratory Data Analysis. Data Science Program.

Visualization

Visualization or **visualization** is any technique for creating images, diagrams, or animations to communicate a message. Visualization through visual imagery has been an effective way to communicate both abstract and concrete ideas since the dawn of humanity.

Examples from history include cave paintings, Egyptian hieroglyphs, Greek geometry, and Leonardo da Vinci's revolutionary methods of technical drawing for engineering and scientific purposes.

A comprehensive list of charts is available at https://datavizcatalogue.com/

Erick Gomez-Nieto, PhD. Exploratory Data Analysis. Data Science Program.

What makes a “good” visualization?

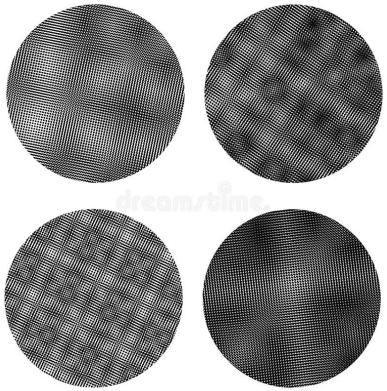
Effective: the viewer gets it (ease of interpretation).

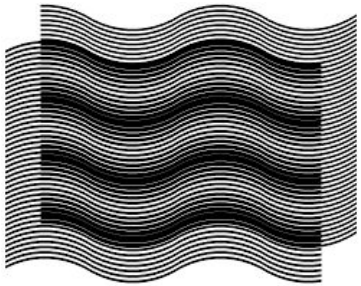
Accurate: sufficient for correct quantitative evaluation. Lie factor = size of visual effect/size of data effect.

Efficient: minimize data-ink ratio and chart-junk, show data, maximize data-ink ratio, brase non-data-ink, brasse redundant data-ink.

Aesthetics: must not offend viewer's senses (e.g. moire patterns).

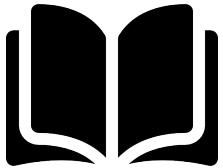
Adaptable: can adjust to serve multiple needs.





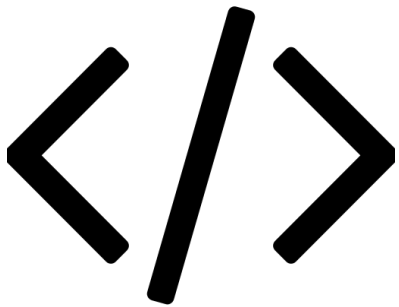
Erick Gomez-Nieto, PhD. Exploratory Data Analysis. Data Science Program.

Erick Gomez-Nieto, PhD. Exploratory Data Analysis. Data Science Program.



What makes a visualization memorable?

Borkin, M. A., Vo, A. A., Bylinskii, Z., Isola, P., Sunkavalli, S., Oliva, A., & Pfister, H. (2013). What makes a visualization memorable?. *IEEE transactions on visualization and computer graphics*, *19*(12), 2306-2315.



So… let’s code!

Erick Gomez-Nieto, PhD. Exploratory Data Analysis. Data Science Program.

Erick Gomez-Nieto, PhD. Exploratory Data Analysis. Data Science Program.

Seaborn

Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

Benefits of using Seaborn

- A dataset-oriented API for examining relationships between multiple variables
- Specialized support for using categorical variables to show observations or aggregate statistics
- Options for visualizing univariate or bivariate distributions and for comparing them between subsets of data
- Automatic estimation and plotting of linear regression models for different kinds dependent variables
- Convenient views onto the overall structure of complex datasets
- High-level abstractions for structuring multi-plot grids that let you easily build complex visualizations
- Concise control over matplotlib figure styling with several built-in themes
- Tools for choosing color palettes that faithfully reveal patterns in your data

Erick Gomez-Nieto, PhD. Exploratory Data Analysis. Data Science Program.

A simple example

```
[128] print(tips)
```

```
      total_bill   tip     sex smoker   day    time  size
0          16.99  1.01  Female     No   Sun  Dinner     2
1          10.34  1.66    Male     No   Sun  Dinner     3
2          21.01  3.50    Male     No   Sun  Dinner     3
3          23.68  3.31    Male     No   Sun  Dinner     2
4          24.59  3.61  Female     No   Sun  Dinner     4
..           ...   ...     ...    ...   ...     ...   ...
239        29.03  5.92    Male     No   Sat  Dinner     3
240        27.18  2.00  Female    Yes   Sat  Dinner     2
241        22.67  2.00    Male    Yes   Sat  Dinner     2
242        17.82  1.75    Male     No   Sat  Dinner     2
243        18.78  3.00  Female     No  Thur  Dinner     2

[244 rows x 7 columns]
```

Erick Gomez-Nieto, PhD. Exploratory Data Analysis. Data Science Program.

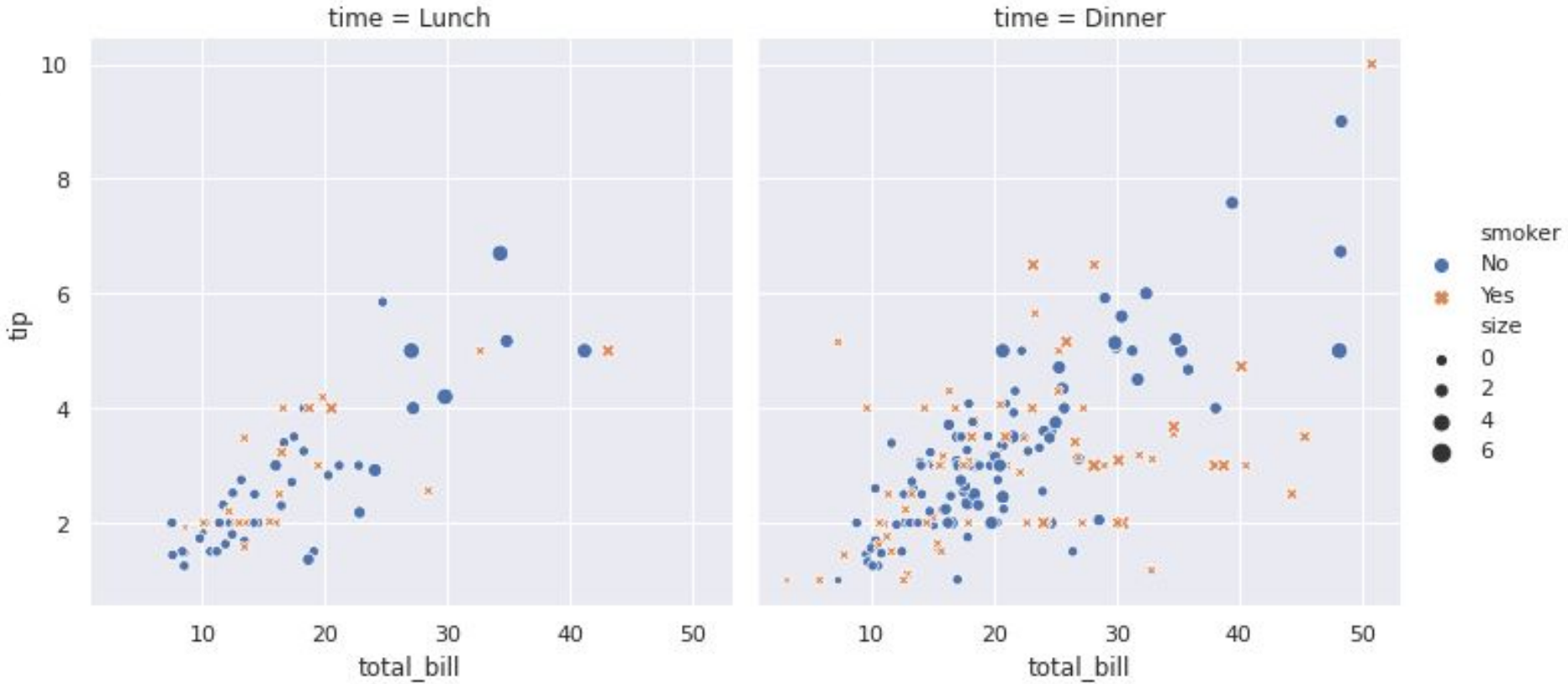
Erick Gomez-Nieto, PhD. Exploratory Data Analysis. Data Science Program.

```
import seaborn as sns  # We import seaborn, which is the only library necessary for this simple example.

sns.set()              # We apply the default default seaborn theme, scaling, and color palette.

tips = sns.load_dataset("tips")    # We load one of the example datasets.

# We draw a faceted scatter plot with multiple semantic variables.
sns.relplot(x="total_bill", y="tip", col="time",
            hue="smoker", style="smoker", size="size",
            data=tips);
```



Erick Gomez-Nieto, PhD. Exploratory Data Analysis. Data Science Program.