UNIVERSITAS
GADJAH MADA

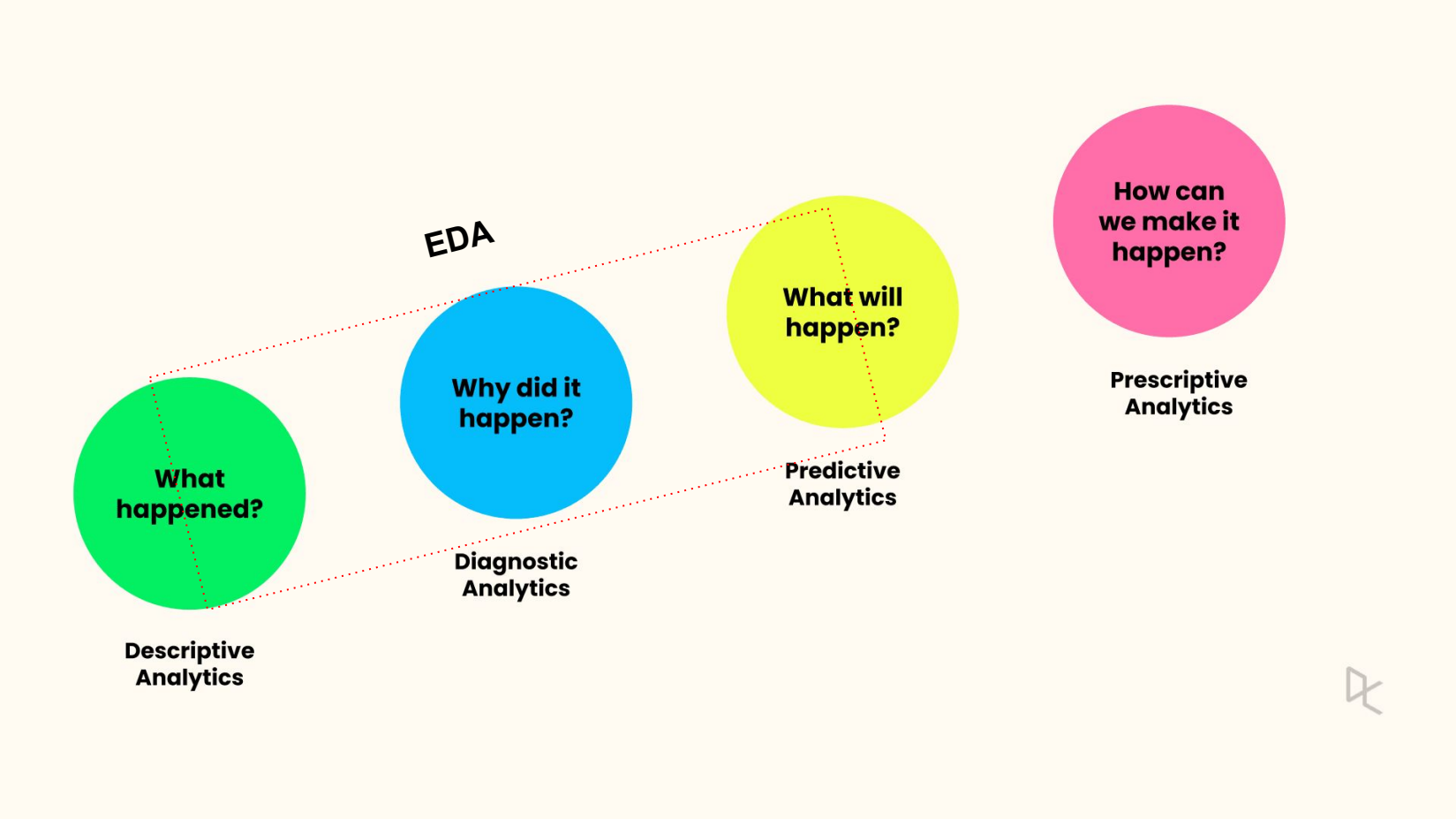
Introduction to **Exploratory Data Analysis (EDA)**

Komatik-Session (Data Mining)

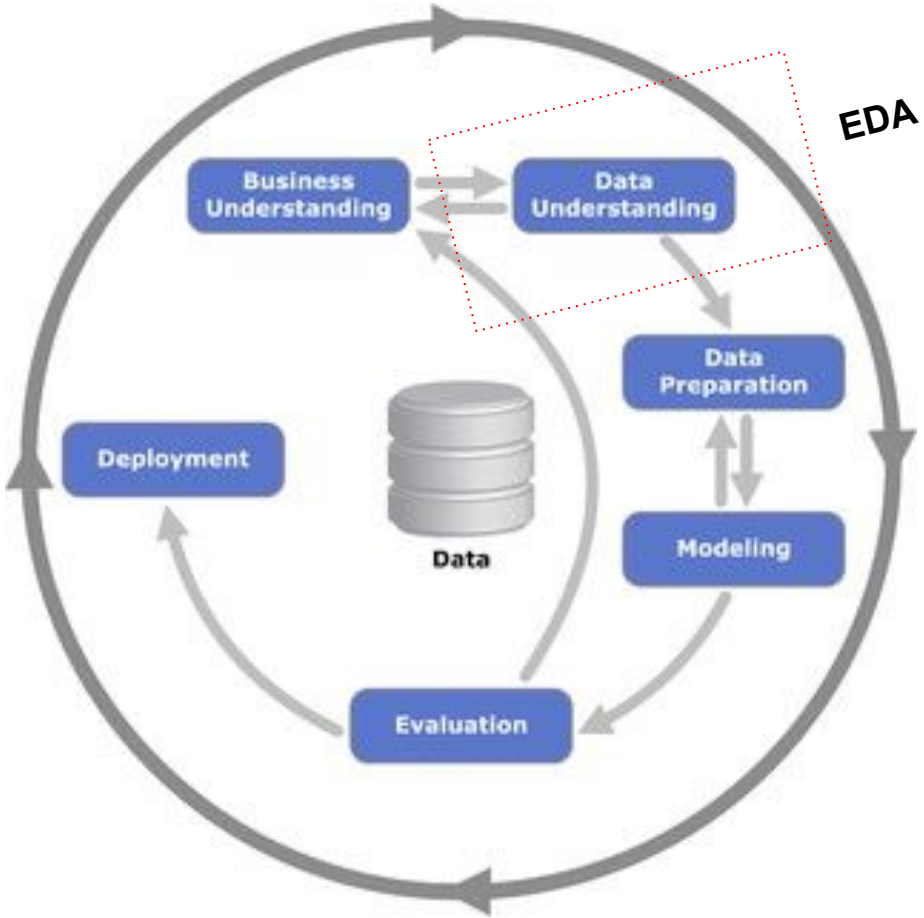
LOCALLY ROOTED,
GLOBALLY RESPECTED

ugm.ac.id

Analytics Stage



CRISP-DM


EDA
Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data

EDA Definition

Exploratory data analysis (EDA) is used by data scientists to **analyze and investigate data sets** and **synthesis their main characteristics,** often employing **data visualization methods**. It helps determine how best to manipulate data sources to get the answers you need, making it easier for data science.

3 Key of EDA

1. Analyze and Investigate
2. Synthesis = Summary + Insight
3. Data visualization

Synthesis vs summary

Summary

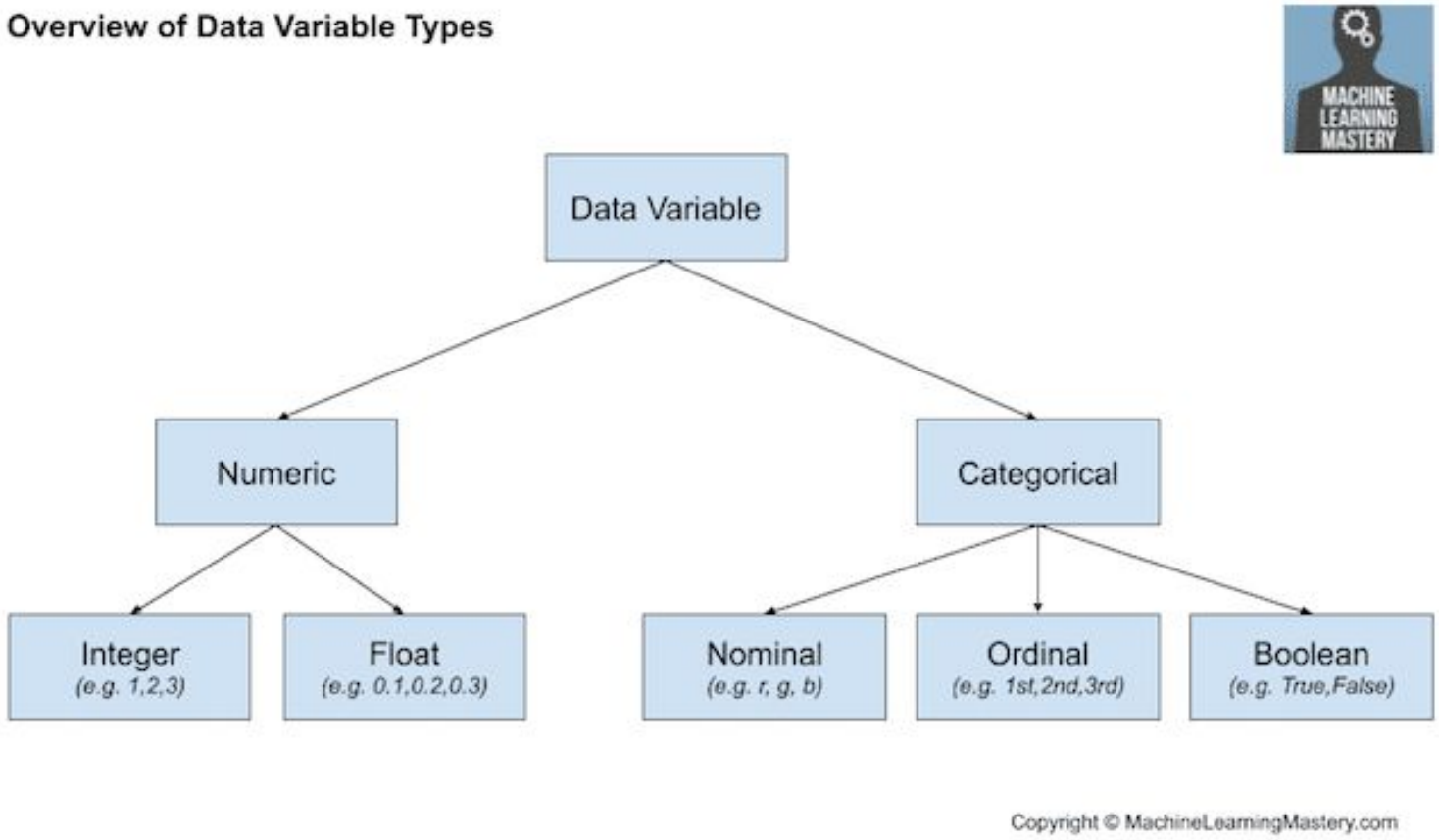
"50% of deals that reached the contract stage typically closed."

Synthesis

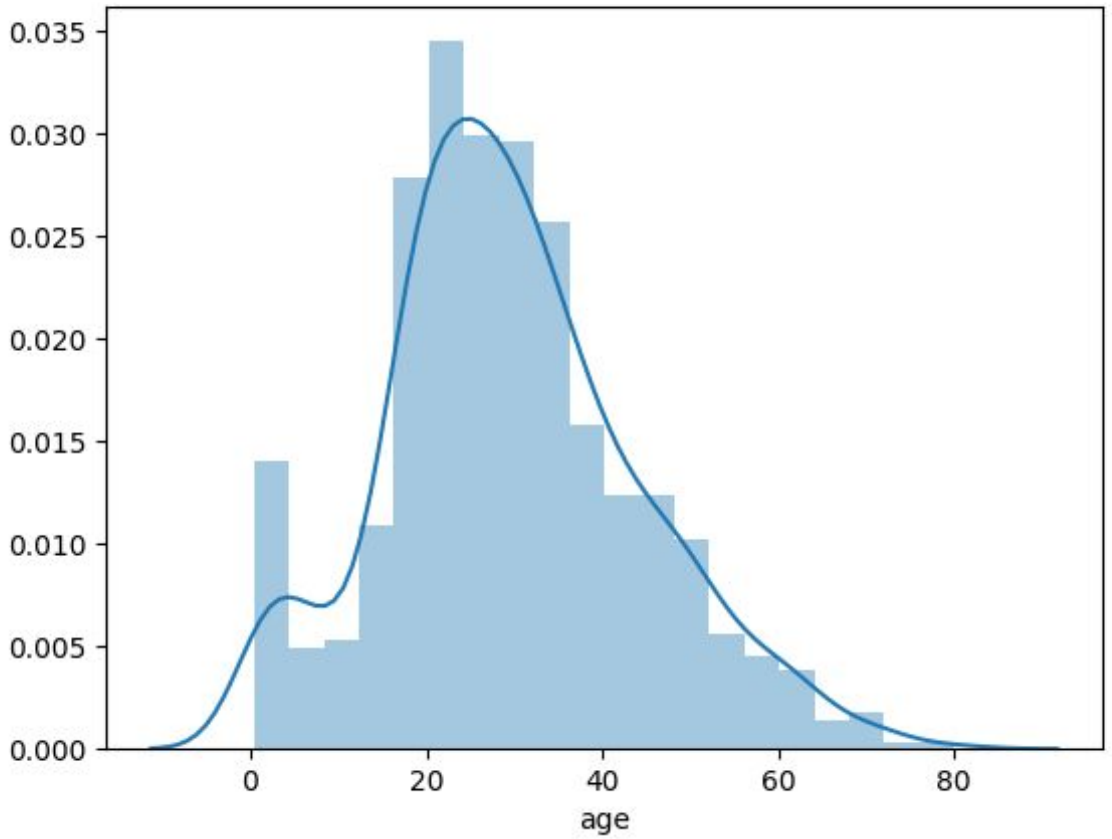
"50% of deals that reach the contract stage typically closed. This is pretty concerning, because it's down from our historical average of 60% close rate. We think there may be two reasons for the decrease: (1) lower sales team productivity, and (2) higher loss rates to our competitor. We need to address this problem quickly, because if we don't, we're going to lose market share quickly."

1. **What's the most important take-away?**
2. **What are the root causes?**
3. **What's the implication?**

Data Variable Type

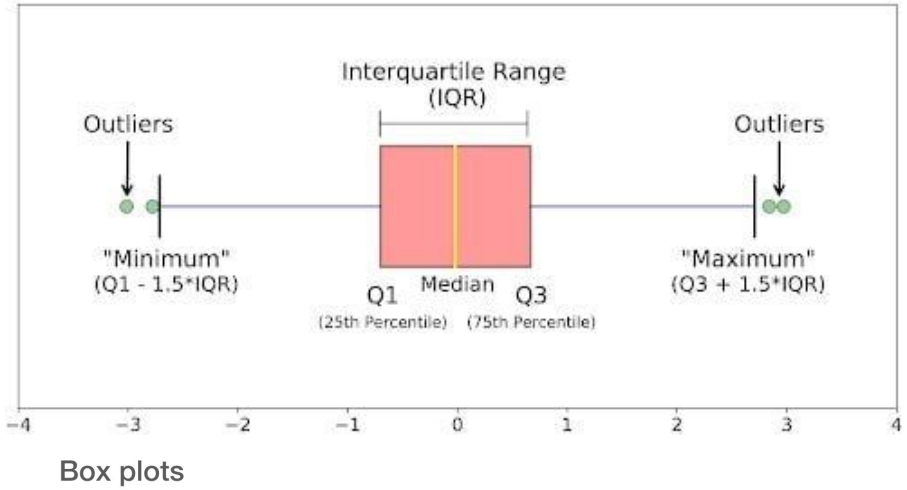


Data Visualization : Distribution Plot

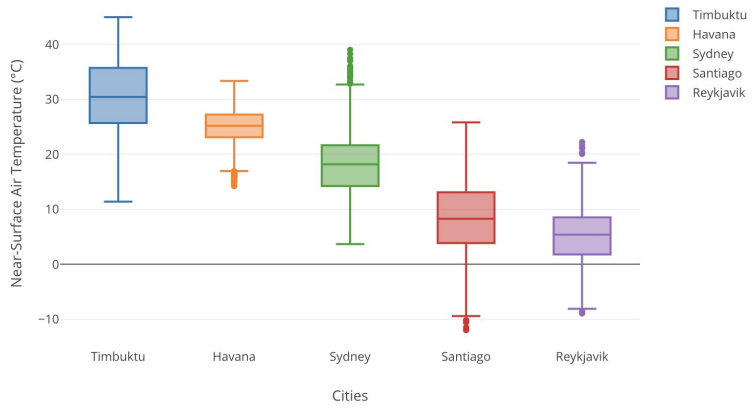


- Numerical-Count/Density

Data Visualization : Box Plot

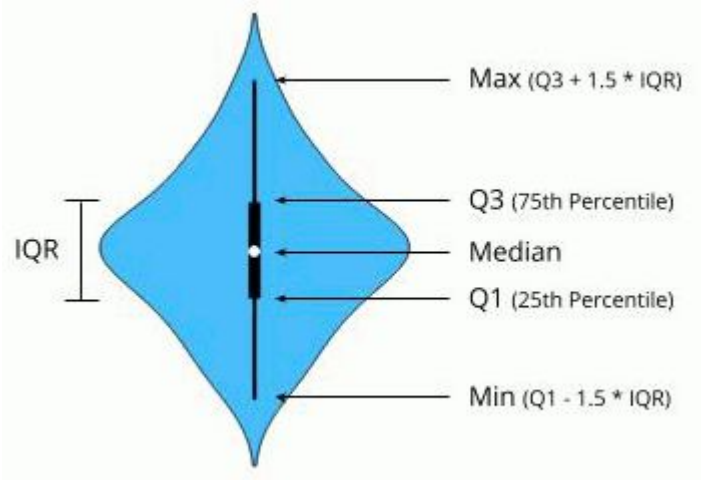


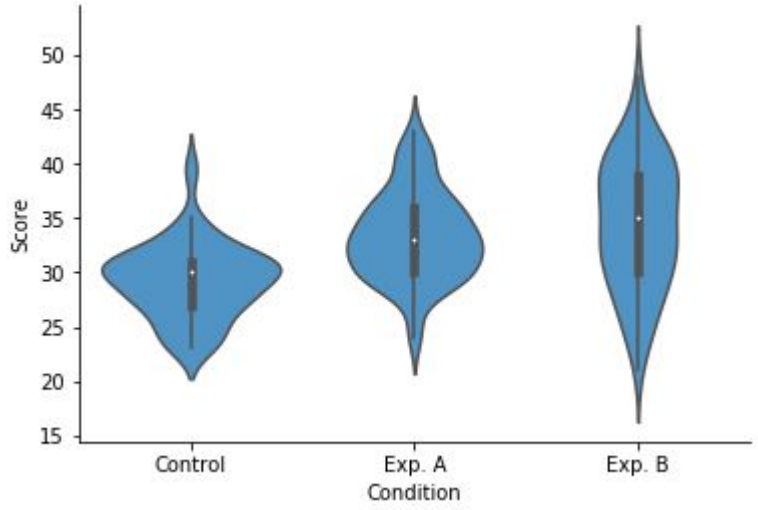
Box plots



- Numerical
- Categorical-Numerical

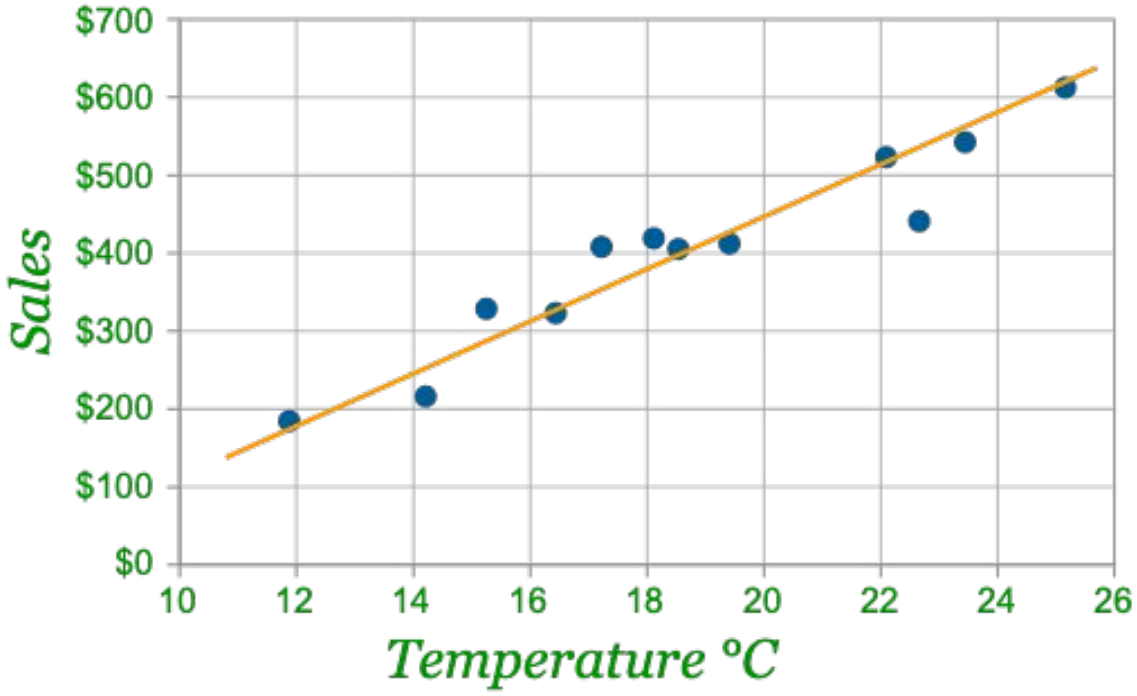
Data Visualization : Violin Plot





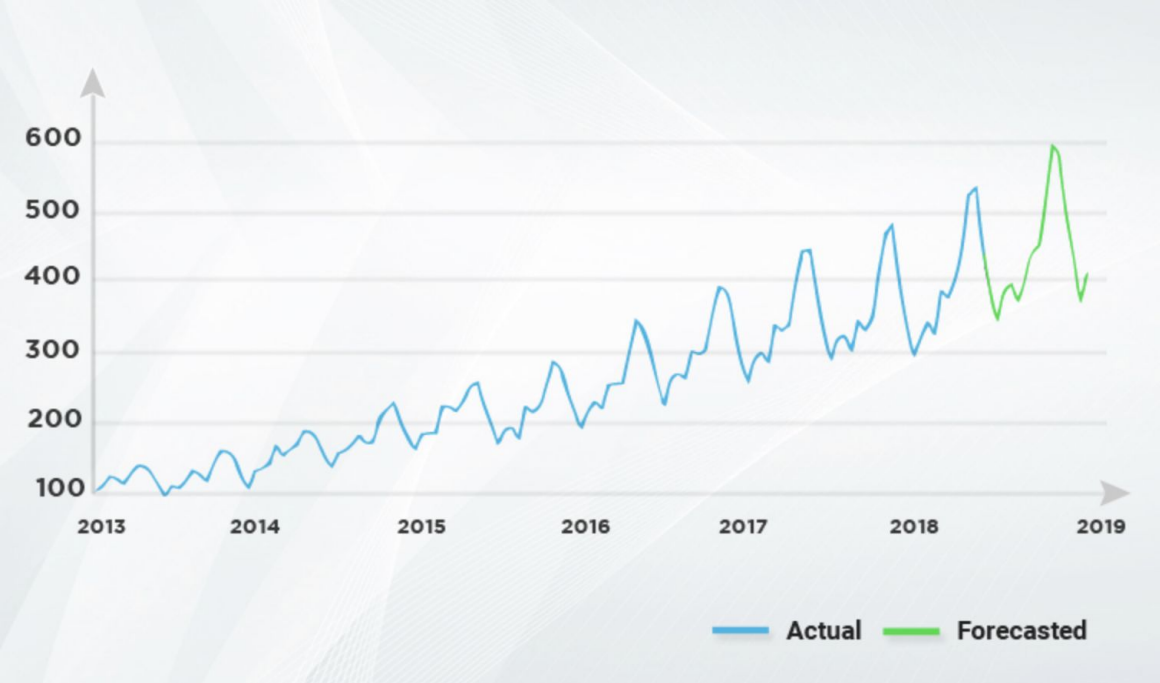
- Numerical
- Categorical-Numerical

Data Visualization : Scatter Plot



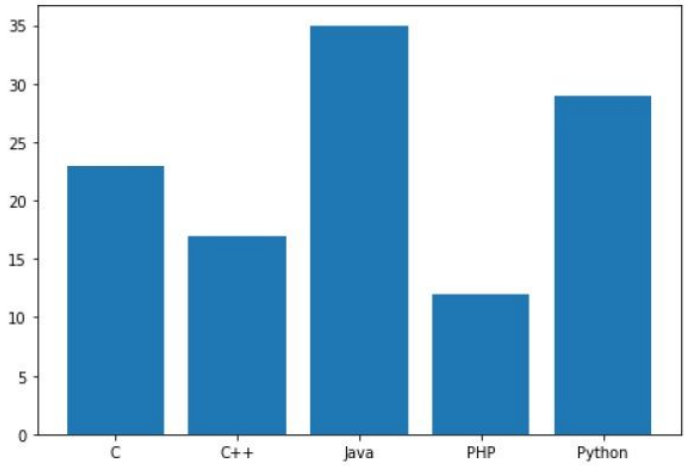
- Numerical-Numerical

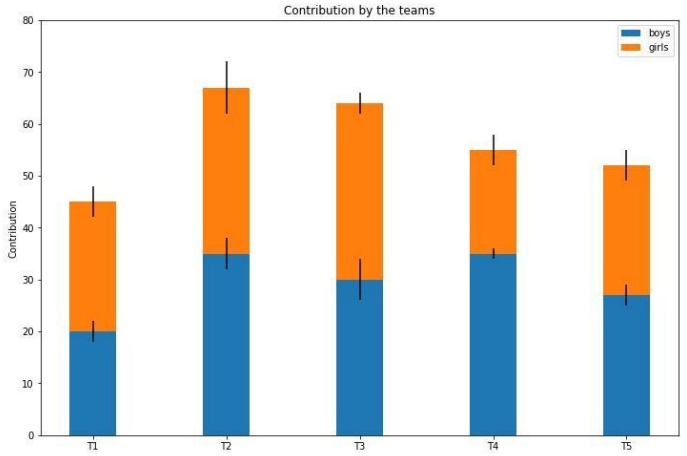
Data Visualization : Line Plot

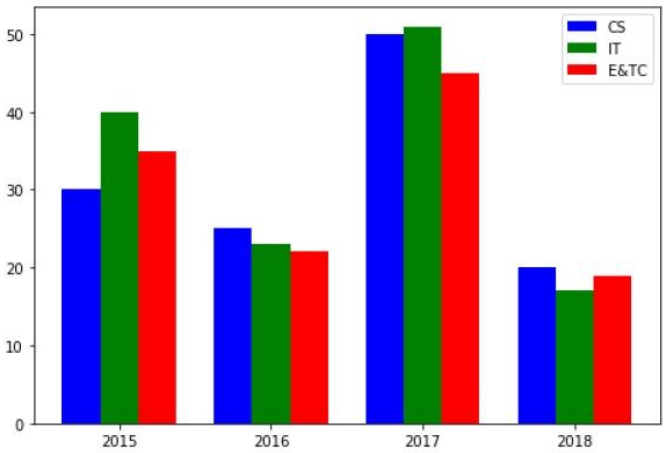


- Numerical -Numerical
- Time Series-Numerical

Data Visualization : Barplot

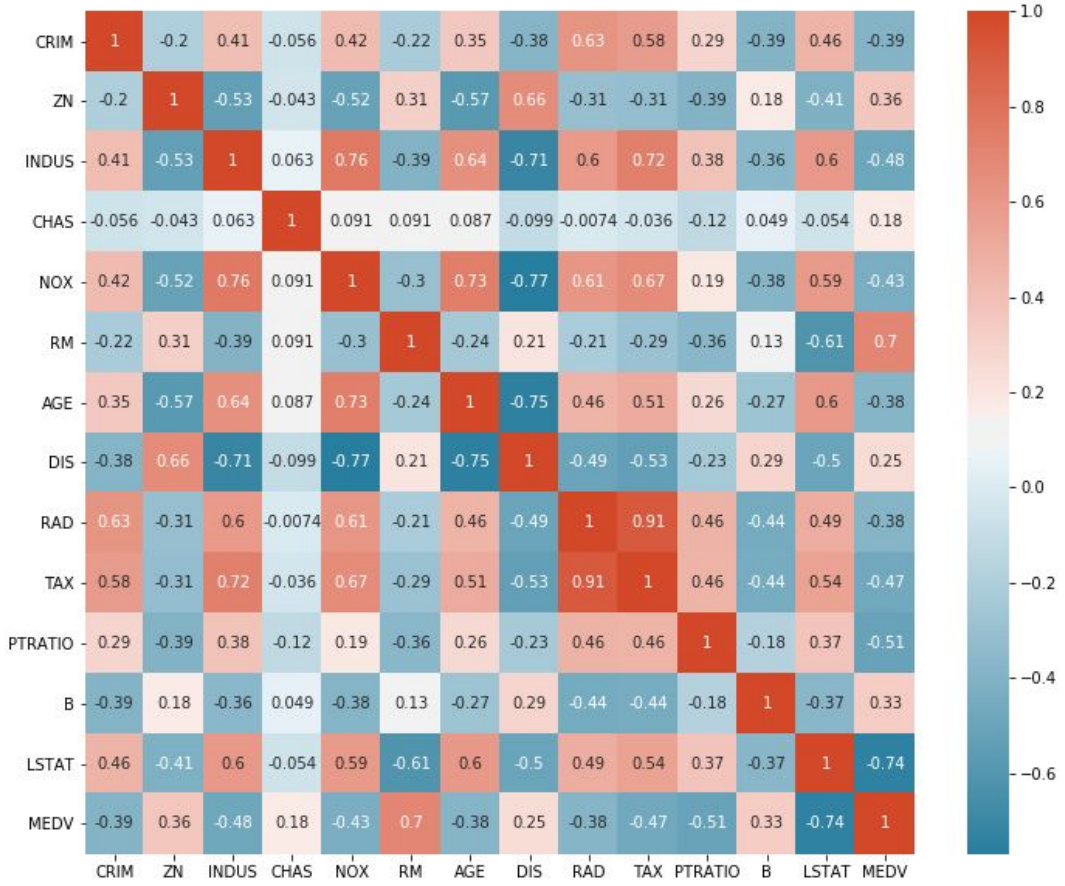






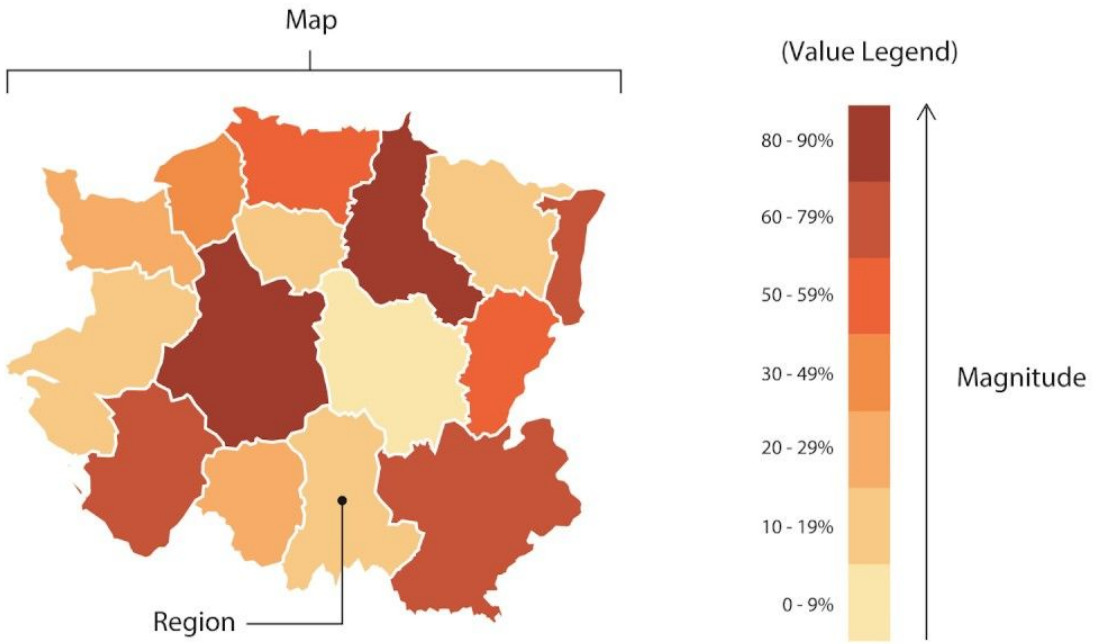
- Categorical-Numerical/ Aggregate

Data Visualization : Heatmap



- Categorical (Variable Name) - Numerical (Correlation)

Data Visualization : Choropleth Map



- Polygon (Region) - Numerical (Correlation)

Tips

1. Set **clear goals**
2. Determine the **target variable** and **focusing** from it
3. Adjust the **data type** with the **visualization**
4. Use aggregate function like **mean, median, mode, count** to **simplify** the data
5. If the **categorical variable** have **many value**, use **top count** in countplot
6. **Deep-dive**, grouping by **categorical variables** if possible
7. Clear **title, axis, legend, label,** or other component
8. Tell your data **cleary** and don't forget to **summarize** [**synthesize** it if possible]
9. **Bold important sentences** or words