Data Science Fundamentals

Exploratory Data Analysis

What is Exploratory Data Analysis (EDA)?

EDA is a process of analyzing data to find insights and patterns that may not be immediately apparent. It involves using statistical techniques and visualization tools to gain insights into the data.

In other words, it's like detective work for data scientists!

EDA is important because it helps us navigate through the maze of data that we are presented with.
By using visualization tools and statistical techniques, we can quickly identify patterns, relationships, and anomalies that would be difficult to find otherwise.

There are several steps involved in EDA, including:

- Data Collection
- Data Cleaning
- Data Exploration
- Data Visualization
- Drawing Conclusions

The first step in EDA is **data collection**. Like a squirrel collecting nuts, we need to gather all the data we can find. It is important to ensure that the data collected is accurate and complete. Remember, garbage in, garbage out!

Data cleaning is the process of identifying and correcting or removing errors, duplicates, and missing values in the data. It's like Marie Kondo-ing your data - only keeping the data that brings you joy.

Data exploration involves summarizing the data using descriptive statistics and visualizations. This helps in identifying outliers, patterns, and relationships within the data. It's like playing hide-and-seek with the data, and our job is to find the patterns!

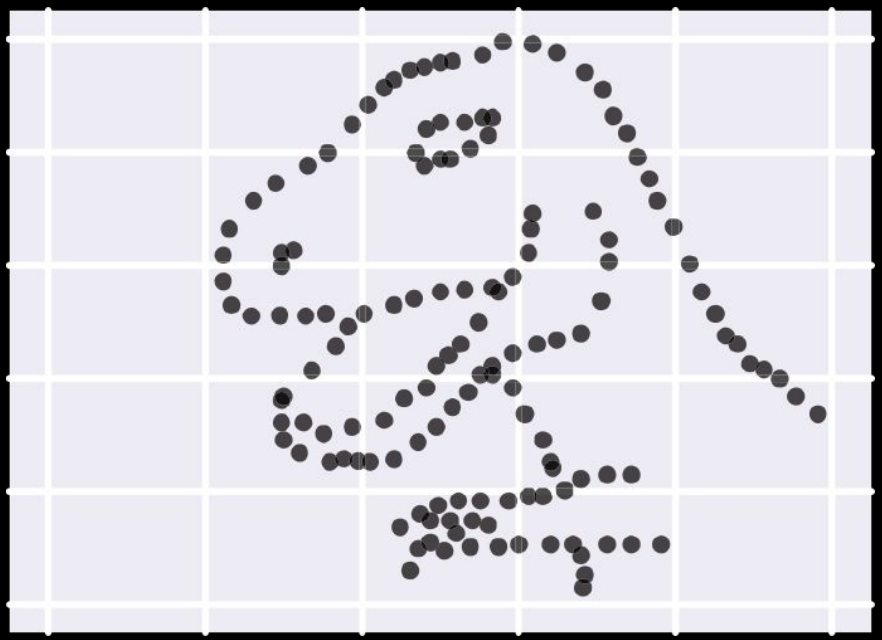
Data visualization is a powerful tool in EDA. It helps in understanding the data better by representing it in a visual form.

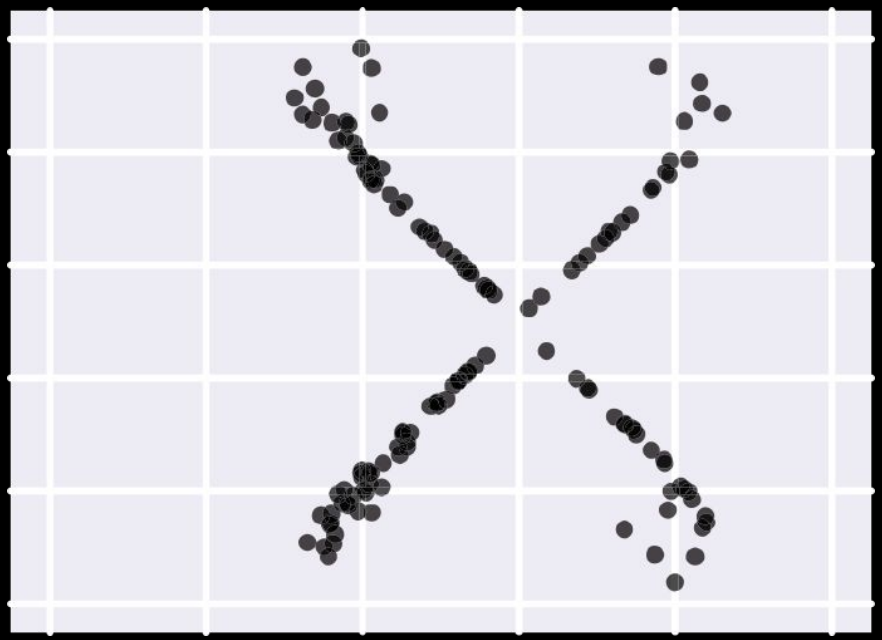
Charts, graphs, and plots are some of the popular visualization tools used in EDA.

It's like a magic show, but instead of rabbits, we're pulling out insights from the data!

Data visualization is crucial, because statistical methods only show a fraction of the data.

The datasets below for example have the same mean and standard deviation, but if you look closely you might spot some subtle differences.





There are several types of visualizations used in EDA, including scatterplots, histograms, boxplots, and heatmaps. Each visualization has its own strengths and weaknesses and is used depending on the type of data being analyzed.

It's like having a toolbox with different tools for different tasks - except we're building insights instead of furniture!

It's dangerous to go alone... that's why there are several tools available for EDA, including **Python, R, Tableau, and Excel.**

These tools have built-in functions and libraries that help in visualizing and analyzing the data.

The final step in EDA is **drawing conclusions** based on the insights gained from the data.

This involves making decisions or taking actions based on the patterns, relationships, and anomalies identified in the data.

Like a detective who solves the case, you will bring justice to the world of data!

To make EDA more effective, here are some tips to keep in mind:

- **Start with a clear understanding of the problem you are trying to solve.**

  It's like looking at a treasure map that leads you to the insights you're looking for!

- Use multiple visualizations to gain different perspectives on the data.
  Just like statistics visualization might only show one aspect of the data.

- Keep the audience in mind when creating visualizations.

  Tailor your presentation to the people who will be using the insights! That also means avoiding fancy jargon sometimes.

- Use statistical tests to confirm the insights gained from visualizations.
  Think of it like a lie detector to make sure the insights are truthful!

- Document your process and findings.
  Yes put out your little diary and write down the jucy data gossip. It will help you remember your data journey

Remember

1. Make sure to collect all relevant data and clean it properly
2. Use multiple visualizations techniques and statistical methods to get a good overview of your data
3. Verify your findings with statistical tests

Feel free to reach out or to connect with me for more weekly slideshows on visualization, data science and machine learning.