2. Exploratory Data Analysis (EDA)

Day 2 of #DataScience28.

Today's subject: Exploratory Data Analysis (#EDA) , a #thread (thread)

#DataScience, #MachineLearning, #66DaysOfData.

Yesterday I covered the topic and importance of Data Cleaning and preparation, if you missed it, you can catch it here: https://jackraiferbaruch.medium.com/day-1-of-datascience28-data-cleaning-and-preparation-8fc8c766350d

Today, I will cover Exploratory Data Analysis (EDA), which is a critical part for the success of any Data Science project. Sometimes referred to as the "playing with the data" phase, here we work to understand the structure, patterns, and relationships in the data, and to identify any important features or anomalies. The goal of EDA is to gain a deeper understanding of the data, to identify any potential issues or limitations, and to inform the later stages of the project.

The importance of EDA cannot be overstated, as it lays the foundation for the rest of the project. Without proper EDA, it is easy to miss important patterns, relationships, or small things in the data that just seem wrong (anomalies), which can lead to incorrect insights, wrong inferences, bad predictions, and unreliable results in the later stages of the project. On the other hand, good EDA can reveal surprising insights and trends, and can change the direction of the project in unexpected ways.

For example, in a predictive modeling project, good EDA might reveal that there is a strong relationship between a particular feature and the target variable. This insight might lead the data team to focus on this feature for their model, which might result in improved predictions and better performance. Similarly, in a customer segmentation project, good EDA might reveal that there are unexpected patterns in the data that can be used to create more meaningful segments, leading to better models.

It is vital to understand that bad EDA can, and often does, break a project by leading the team down a wrong path or an impossible rabbit hole. For example, sloppy EDA may suggest that there is a relationship between two variables when, in fact, there is not. This can lead to including the wrong features in a model, which will probably result in incorrect predictions and poor performance. Similarly, if the EDA is performed incompletely, we might miss important patterns or anomalies, which can lead to incorrect conclusions.

In conclusion, Exploratory Data Analysis (EDA) is critical to the success of any Data Science project, either leading it in the wrong direction, derailing the efforts of the data team, or helping us discover new and interesting ways to create value with our data. Therefore, it's important for any data scientists to take the time and perform a thorough and comprehensive EDA, to ensure that the data is understood, find good insights, and create good inferences that will increase the chances of creating a successful and valuable model.