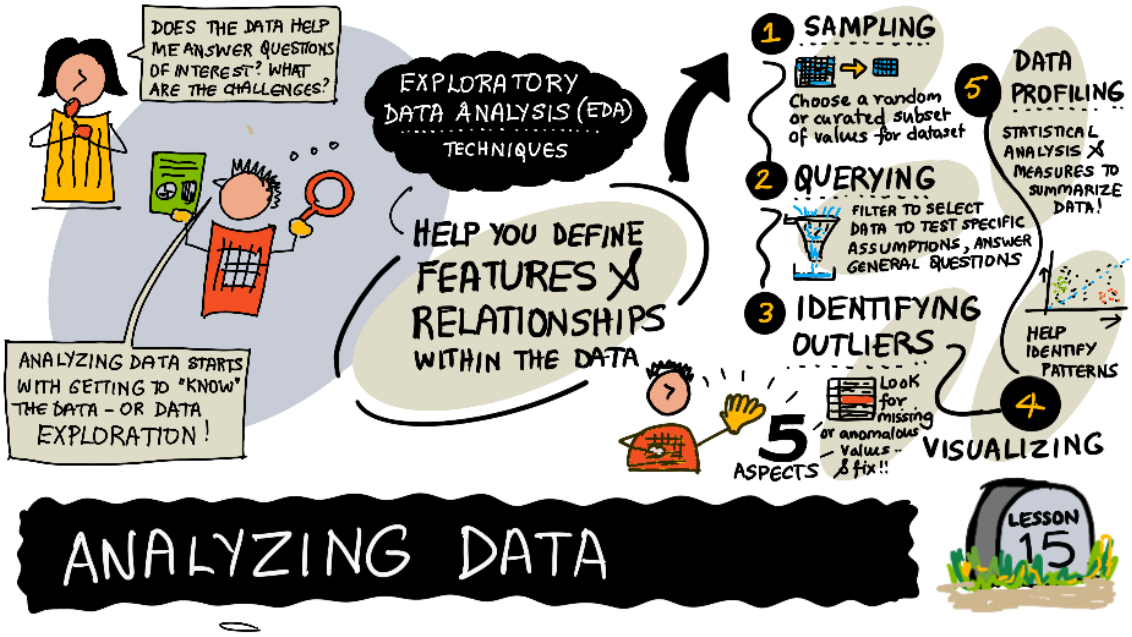
The Data Science Lifecycle: Analyzing



Data Science Lifecycle: Analyzing - *Sketchnote by @nitya*

Pre-Lecture Quiz

Pre-Lecture Quiz

Analyzing in the data lifecycle confirms that the data can answer the questions that are proposed or solving a particular problem. This step can also focus on confirming a model is correctly addressing these questions and problems. This lesson is focused on Exploratory Data Analysis or EDA, which are techniques for defining features and relationships within the data and can be used to prepare the data for modeling.

We'll be using an example dataset from Kaggle to show how this can be applied with Python and the Pandas library. This dataset contains a count of some common words found in emails, the sources of these emails are anonymous. Use the notebook in this directory to follow along.

Exploratory Data Analysis

The capture phase of the lifecycle is where the data is acquired as well as the problems and questions at hand, but how do we know the data can help support the end result? Recall that a data scientist may ask the following questions when they acquire the data:

- Do I have enough data to solve this problem?
- Is the data of acceptable quality for this problem?
- If I discover additional information through this data, should we consider changing or redefining the goals?

Exploratory Data Analysis is the process of getting to know that data and can be used to answer these questions, as well identify the challenges of working with the dataset. Let's focus on some of the techniques

used to achieve this.

Data Profiling, Descriptive Statistics, and Pandas

How do we evaluate if we have enough data to solve this problem? Data profiling can summarize and gather some general overall information about our dataset through techniques of descriptive statistics. Data profiling helps us understand what is available to us, and descriptive statistics helps us understand how many things are available to us.

In a few of the previous lessons, we have used Pandas to provide some descriptive statistics with the `describe()` function. It provides the count, max and min values, mean, standard deviation and quantiles on the numerical data. Using descriptive statistics like the `describe()` function can help you assess how much you have and if you need more.

Sampling and Querying

Exploring everything in a large dataset can be very time consuming and a task that's usually left up to a computer to do. However, sampling is a helpful tool in understanding of the data and allows us to have a better understanding of what's in the dataset and what it represents. With a sample, you can apply probability and statistics to come to some general conclusions about your data. While there's no defined rule on how much data you should sample it's important to note that the more data you sample, the more precise of a generalization you can make of about data. Pandas has the `sample()` function in its library where you can pass an argument of how many random samples you'd like to receive and use.

General querying of the data can help you answer some general questions and theories you may have. In contrast to sampling, queries allow you to have control and focus on specific parts of the data you have questions about. The `query()` function in the Pandas library allows you to select columns and receive simple answers about the data through the rows retrieved.

Exploring with Visualizations

You don't have to wait until the data is thoroughly cleaned and analyzed to start creating visualizations. In fact, having a visual representation while exploring can help identify patterns, relationships, and problems in the data. Furthermore, visualizations provide a means of communication with those who are not involved with managing the data and can be an opportunity to share and clarify additional questions that were not addressed in the capture stage. Refer to the section on Visualizations to learn more about some popular ways to explore visually.

Exploring to identify inconsistencies

All the topics in this lesson can help identify missing or inconsistent values, but Pandas provides functions to check for some of these. isna() or isnull() can check for missing values. One important piece of exploring for these values within your data is to explore why they ended up that way in the first place. This can help you decide on what actions to take to resolve them.

Pre-Lecture Quiz

Assignment

Exploring for answers