EDA with Python

Overview

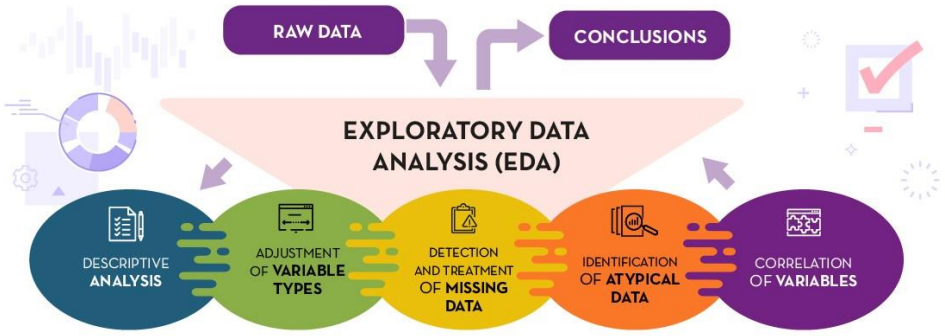
Knowledge of the data that we will be working with is a key step in the analysis and, let's face it, in practically every endeavor that entails utilizing data in some way, whether by mapping it out or presenting it in a graphical form. Because of this, exploratory data analysis (EDA) is a crucial procedure that entails running preliminary analyses on the data to find patterns, spot anomalies, test hypotheses, and confirm presumptions with the aid of graphical and statistical representations of the data. This session takes you through the different techniques that are followed when performing EDA in Python.

Outcomes

At the end of this session, you will be able to:

- Describe the importance of EDA and its necessity in data analysis
- Perform data cleaning, data transformation, data imputation, and outlier treatment in Python
- Apply statistical concepts such as bivariate analysis in Python

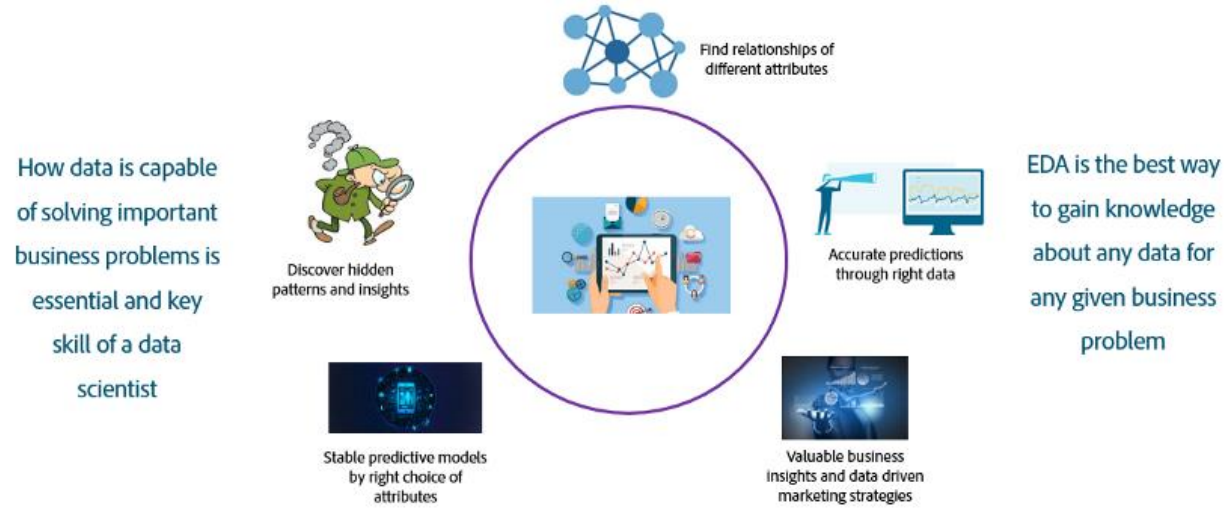
What is EDA?



Source

Exploratory data analysis, or EDA for short, is all about exploring your data to see what you can discover about it, what patterns you can spot, and what connections you can make. EDA is a crucial first step in the analysis and model-building processes. It almost always aids in revealing parts of your data that you wouldn't have otherwise noticed. When done properly, it can also help you design new queries and subjects for research. To get an understanding of the goals of EDA and the general EDA techniques, refer to the following link:
https://www.codecademy.com/courses/eda-exploratory-data-analysis-python/articles/what-is-eda

Why is EDA important?



Source

EDA: Inspect, clean, and validate a dataset

Finding the root causes of data problems and determining the best way to solve them is one of the most difficult aspects of data cleaning. Exploratory data analysis (EDA) is a very effective tool for doing this. It is crucial to keep in mind that each dataset is unique and will necessitate a different kind of investigation. Following the data, checking your presumptions, and looking into anything unexpected are the three pillars of EDA. The purpose of the link below is to show how EDA may guide the initial data inspection, address questions based on the obtained information, inspect missing data, and also provides a video on data exploration:
https://www.codecademy.com/courses/eda-exploratory-data-analysis-python/articles/eda-inspect-clean-and-validate-a-dataset

EDA Project: Diagnosing Diabetes

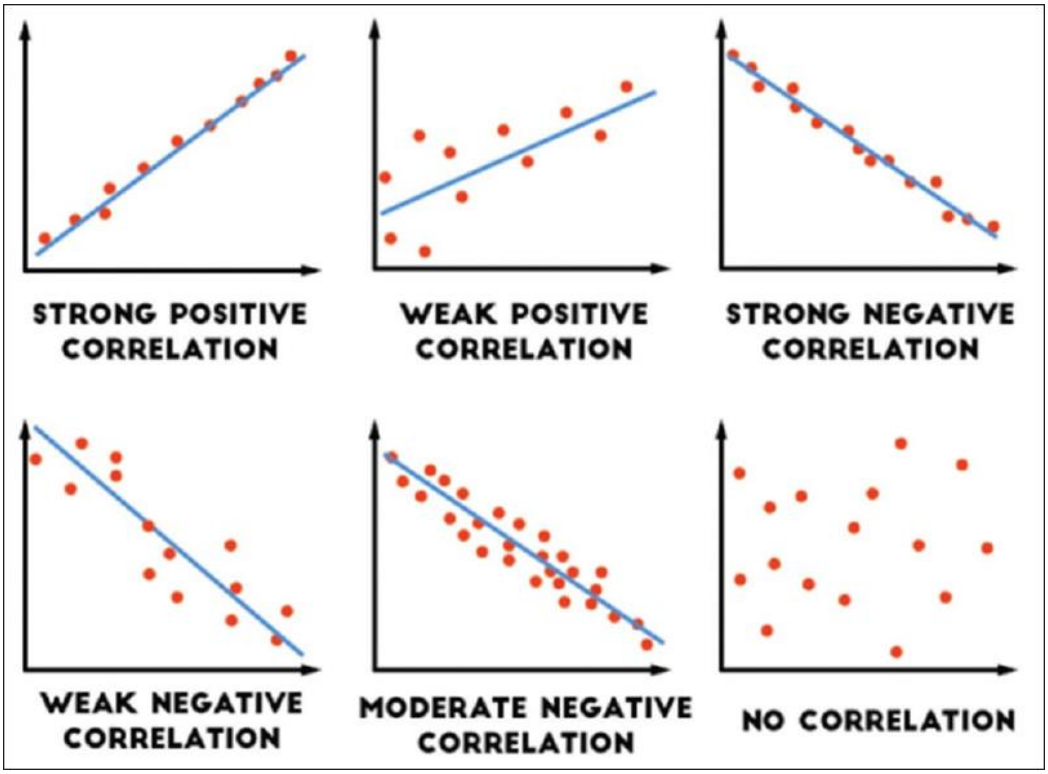
Dive into a complete hands-on EDA project where we are exploring data that looks at how certain diagnostic factors affect the diabetes outcome of women patients. The dataset is from the National Institute of Diabetes and Digestive and Kidney Diseases.
You will get your hands dirty with familiarizing yourself with the data, and the shape of the data, inspecting and dealing with missing values, displaying the summary statistics, and spot outliers and check for incorrect data types of columns in the following link:
(if you are stuck, use the 'Get a hint' option for help)
https://www.codecademy.com/courses/eda-exploratory-data-analysis-python/projects/data-cleaning-project

Summarizing a Single Feature

It is frequently beneficial to conduct some preliminary explorations of the data using exploratory data analysis (EDA) to have a better understanding of what you will be dealing with before entering into a formal analysis with a dataset. Basic summary statistics and visualizations are crucial elements of EDA since they help us reduce a big volume of data to a manageable number of numbers or visuals.

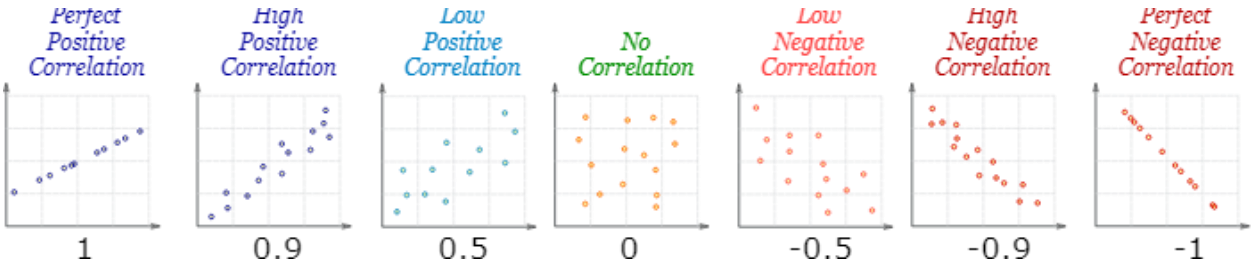
The following link focuses on univariate analysis where each variable is examined independently. This helps respond to inquiries regarding each distinct feature. To get started on summary statistics of both quantitative and qualitative variables, check out the following link:
(all the lessons under 'Data Summaries'. If you are stuck, use the 'Get a hint' option for help)
https://www.codecademy.com/courses/eda-exploratory-data-analysis-python/lessons/data-summaries/exercises/introduction

Introducing Correlation



Source

Any dataset we want to study contains several fields (i.e., columns) of numerous observations (i.e., variables) that reflect various facts. Since they were gathered from the same event, the columns in a dataset are almost always connected. A field of a record may or may not have an impact on another field's value. We must look for dependencies among variables to study the kind of links these columns have and to assess the causes and effects between them. Correlation is a term used to describe the strength of a link between two fields in a dataset, and it is represented by a number between -1 and 1.



Source: https://www.mathsisfun.com/data/correlation.html

To get a quick introduction to correlation and how it works, refer:
https://learning.oreilly.com/library/view/hands-on-exploratory-data/9781789537253/8e2546e8-cea8-4233-a70a-2e5422e12f68.xhtml

Bivariate Analysis

To determine whether there is a relationship between two different variables, bivariate analysis is utilized. Bivariate analysis typically aids us in predicting a value for one variable (i.e., a dependent variable) if we are aware of the value of the independent variable. To get a brief introduction, along with a few practice examples of bivariate analysis, refer to:
https://learning.oreilly.com/library/view/hands-on-exploratory-data/9781789537253/4f8a0095-b73f-4146-b3d4-4e7f3f80c9b7.xhtml

So far, we learned about the various aspects of EDA which will help you to participate actively in the live session. In the live session, you will learn to apply different techniques that are followed when performing EDA with Python.