Print

Introduction to Exploratory Data Analysis

In the first lesson in this module, Describing Data, you were introduced to descriptive statistics.

You learned that you can use numerical summaries and graphical displays to describe the shape, centering, and spread of a distribution.

You also learned that you can use descriptive tools to summarize or visualize two or more variables at a time.

You saw that the type of graph or visualization that you should use depends on the type of data and the number of variables you are analyzing.

For continuous data, the basic visual tools are histograms, box plots, run charts, and scatterplots.

For categorical data, bar charts and mosaic plots are often used.

You also learned how to use dynamically linked graphs to explore potential relationships between variables.

The process of using numerical summaries and visualizations to explore your data and identify potential relationships between variables is called Exploratory Data Analysis, or EDA.

Exploratory data analysis is an investigative process in which you use summary statistics and graphical tools to get to know your data and understand what you can learn from it.

With EDA, you can find anomalies in your data, such as outliers or unusual observations, uncover patterns in your data, understand potential relationships between variables, and generate interesting questions or hypotheses that you'll test using more formal statistical methods.

Exploratory data analysis is like detective work. You're searching for clues and insights that can lead to the identification of potential root causes of the problem you are trying to solve. You explore one variable at a time, then two variables at a time, and then many variables at a time.

You use a variety of graphs and exploratory tools, and you go where your data take you. If one graph or analysis isn't informative, you look at the data from another perspective.

In contrast, many of the traditional statistical methods you'll learn later in the course are used for Confirmatory Data Analysis, or CDA. Tools such as hypothesis testing and explanatory modeling, which build on what you learned through exploratory data analysis, enable you to make formal decisions and draw inferences from your data.

In this lesson, you learn how to create a broader range of visualizations for exploratory data analysis, including visualizations for more complex data that are multidimensional in nature.

You also learn how to stratify your data, how to explore "what if" scenarios, and how to efficiently explore many variables at a time.

This will be somewhat of a survey of exploratory methods. You might not use all of the methods, but you'll learn which methods are available and when they might be useful.

Note that, in this lesson, we focus on the use of visualization where the goal is data exploration and discovery.

In the next lesson, you learn how to use graphical displays and interactive visualizations for data presentation, where your goal is to effectively communicate your findings to stakeholders.

Close