EDA with Python

With the end of this session, you should now be able to:

- Describe the general procedure of EDA in data analysis using Python
- Perform data cleaning tasks such as filtering, dropping, and new variable creation
- Perform data transformation techniques to bin and merge data
- Create single and multivariable data summaries using cross-tab, group-by, and pivot tables
- Perform statistical analysis such as bivariate analysis and obtain correlation graphs and measures between the variables in the dataset
- Perform imputation techniques to handle missing values in the data
- Perform outlier treatment to handle abnormalities in the data

Handling Missing Data

Real-world data isn't always clean and uniform, which is how it differs from the data used in many tutorials. Particularly, many intriguing datasets will contain some missing data. The fact that different data sources may identify missing data in various ways further complicates the situation.

To get a refresher on some general missing data considerations, how Pandas chooses to represent missing data, and examples of various built-in Pandas utilities for handling missing data in Python, go ahead and click on:

https://learning.oreilly.com/library/view/python-data-science/9781098121211/ch16.html

Handling Outliers

In a data sample, outliers are typically described as observations that are abnormally distant from other observations. They represent unusual values in a dataset, to put it another way. Naturally, the abnormal distance we're referring to has no standard measurement and is entirely dependent on the dataset you're looking at. Based on their expertise and functional knowledge of the business reality represented by the dataset, the analyst will determine the distance beyond which to deem other distances abnormal.

To brush up on your knowledge about what outliers are, what causes these outliers in the dataset, widely used approaches in dealing with them, identifying univariate and multivariate outliers, and an example of how to implement outlier detection in Python, refer to:

(only subtopics 'What outliers are and how to deal with them', 'Identifying outliers' and 'Implementing outlier detection in Python')

https://learning.oreilly.com/library/view/extending-power-bi/9781801078207/B17081_12_Final_NM_ePub.xhtml#:-:text=What%20outliers%20are%20and%20how%20to%20deal%20with%20them

Summarizing the Relationship Between Two Features

Univariate data is used for conducting studies that focus on a single variable. For instance, you might research a group of university students to learn their typical SAT scores or a group of diabetic patients to learn their weights. When two variables are being studied, you have bivariate data. For instance, if you are researching a group of college students to determine their typical SAT score and age, you need to uncover two parts of the puzzle (SAT score and age).

There are numerous real-world applications for bivariate data. For instance, knowing when a natural event might happen is quite useful. Bivariate data analysis is one of the most important tools in the

statistician's toolbox. Sometimes all it takes to understand what the data is trying to tell you is to plot one variable against another on a Cartesian plane.
If you are interested to learn more about bivariate analysis and practice on the different techniques involved, check out the following links.

QUANTITATIVE AND CATEGORICAL VARIABLES
The following link discusses methods for evaluating the relationship between a quantitative variable and a categorical variable using statistical differences, side-by-side boxplots, overlapping histograms, and exploring non-binary categorical variables:
(all lessons under 'ASSOCIATIONS: QUANTITATIVE AND CATEGORICAL VARIABLES')
https://www.codecademy.com/courses/eda-exploratory-data-analysis-python/lessons/associations-quantitative-and-categorical-variables/exercises/introduction-to-quantitative-and-categorical-variable-association

TWO QUANTITATIVE VARIABLES
The following link discusses methods for evaluating the relationship between two quantitative variables using scatter plots, exploring covariance and correlation:
(all lessons under 'ASSOCIATIONS: TWO QUANTITATIVE VARIABLES')
https://www.codecademy.com/courses/eda-exploratory-data-analysis-python/lessons/analyzing-associations-two-quantitative-variables/exercises/analyzing-associations-two-quantitative-introduction

TWO CATEGORICAL VARIABLES
The following link discusses methods for evaluating the relationship between two categorical variables using frequency tables, proportion tables, marginal proportions, expected contingency tables, and the chi-square statistic:
(all lessons under 'ASSOCIATIONS: TWO CATEGORICAL VARIABLES')
https://www.codecademy.com/courses/eda-exploratory-data-analysis-python/lessons/associations-two-categorical-variables/exercises/introduction

With this, we reached the end of the session.