# [Exploratory Data Analysis (Module 3)](https://www.coursera.org/learn/data-analysis-with-python/home/module/3)

Table of Contents

[Data Wrangling (Module 2) 1](#_Toc187337790)

[Lesson Summary 1](#_Toc187337791)

### Lesson Summary

Congratulations! You have completed this lesson. At this point in the course, you know:

* Tools like the **'describe'** function in pandas can quickly calculate key statistical measures like mean, standard deviation, and quartiles for all numerical variables in your data frame.
* Use the **'value\_counts'** function to summarize data into different categories for categorical data.
* Box plots offer a more visual representation of the data's distribution for numerical data, indicating features like the median, quartiles, and outliers.
* Scatter plots are excellent for exploring relationships between continuous variables, like engine size and price, in a car data set.
* Use Pandas' **'groupby'** method to explore relationships between categorical variables.
* Use pivot tables and heat maps for better data visualizations.
* Correlation between variables is a statistical measure that indicates how the changes in one variable might be associated with changes in another variable.
* When exploring correlation, use scatter plots combined with a regression line to visualize relationships between variables.
* Visualization functions like **regplot,** from the **seaborn** library, are especially useful for exploring correlation.
* The **Pearson correlation**, a key method for assessing the correlation between continuous numerical variables, provides two critical values—the coefficient, which indicates the strength and direction of the correlation, and the P-value, which assesses the certainty of the correlation.
* A correlation coefficient close to 1 or -1 indicates a strong positive or negative correlation, respectively, while one close to zero suggests no correlation.
* For P-values, values less than .001 indicate strong certainty in the correlation, while larger values indicate less certainty. Both the coefficient and P-value are important for confirming a strong correlation.
* Heatmaps provide a comprehensive visual summary of the strength and direction of correlations among multiple variables.