Week 3

Exploratory Data Analysis

EDA is the prelimonary step in data analysis.

It is used to:

1. Summarize main characteristics of the data
2. Gain better understanding of the data set.
3. Uncover relationships between variales
4. Extract important variables

This module helps us to understand which characteristics have the most impact on the result column.

Learning Objectives:

Descriptive Statistics

Groupby

ANOVA - ANOVA, the analysis of variance a statistical method in which the variation

in a set of observations is divided into distinct components

Correlation

Correlation -Statistics (pearsing correlation and correation heatmapp)

Descriptive statistics

need to understand the data before modelling.

this can be done by descriptive statistics

Descriptive statistics describes basic features of data.

Giving short summaries about the sample and measure of the data.

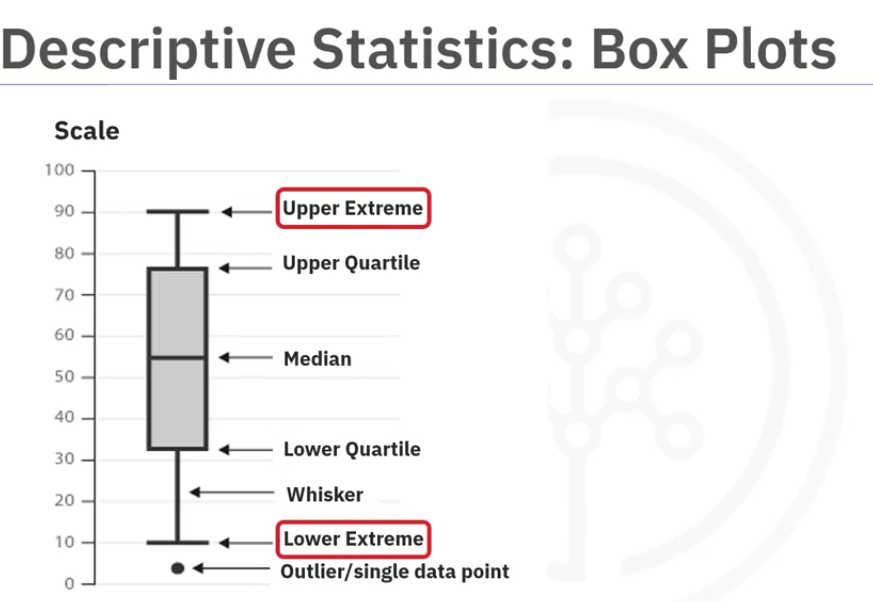
Methods:

df.describe()

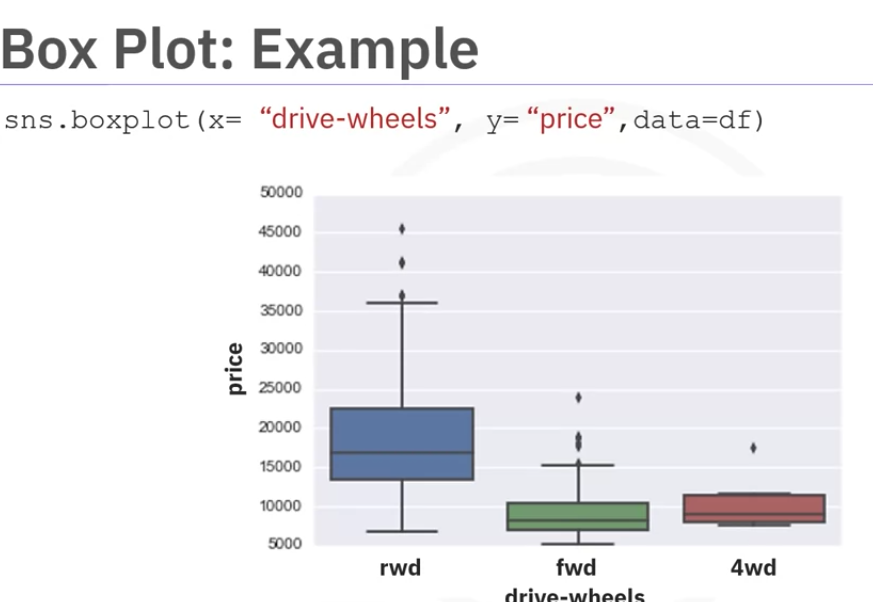
df[‘column\_name’].value\_counts()

Graphs for descriptive statistics:

Box plots



sns.boxpot(x= ‘x label’, y= ‘y-labe, data = df)



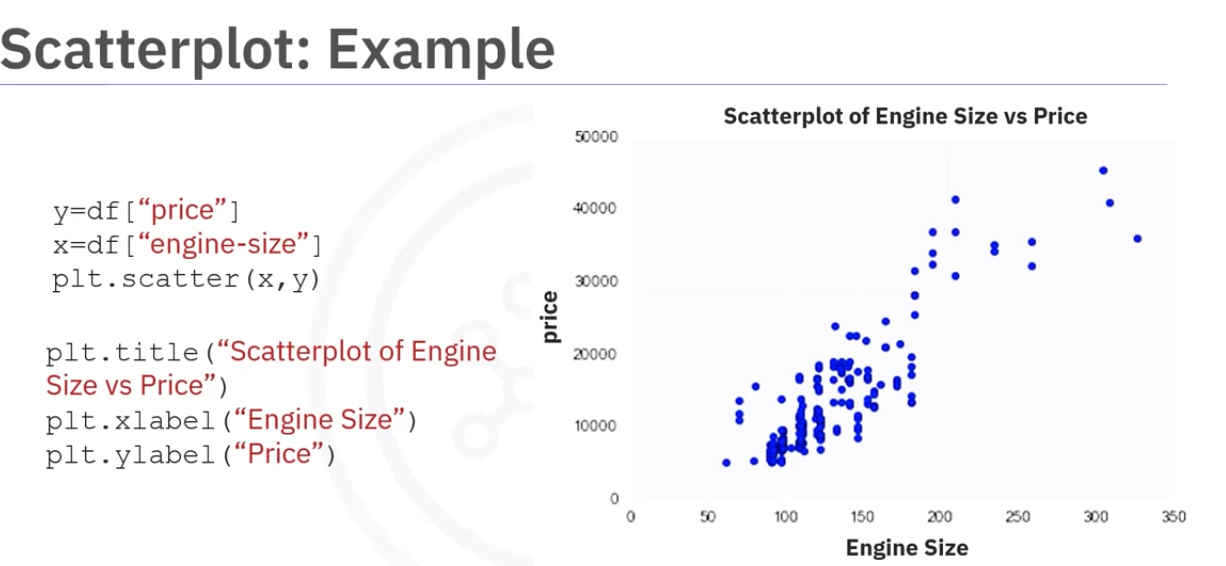
Scatter Plot:

Normally used for Continuous variables.

Example the effect of Size of the engine on price

Needs 2 variables

1. Predictor/ independent variabe on x axis (engine size)
2. Target/ dependent variable on y axis(Price)



Group By

.

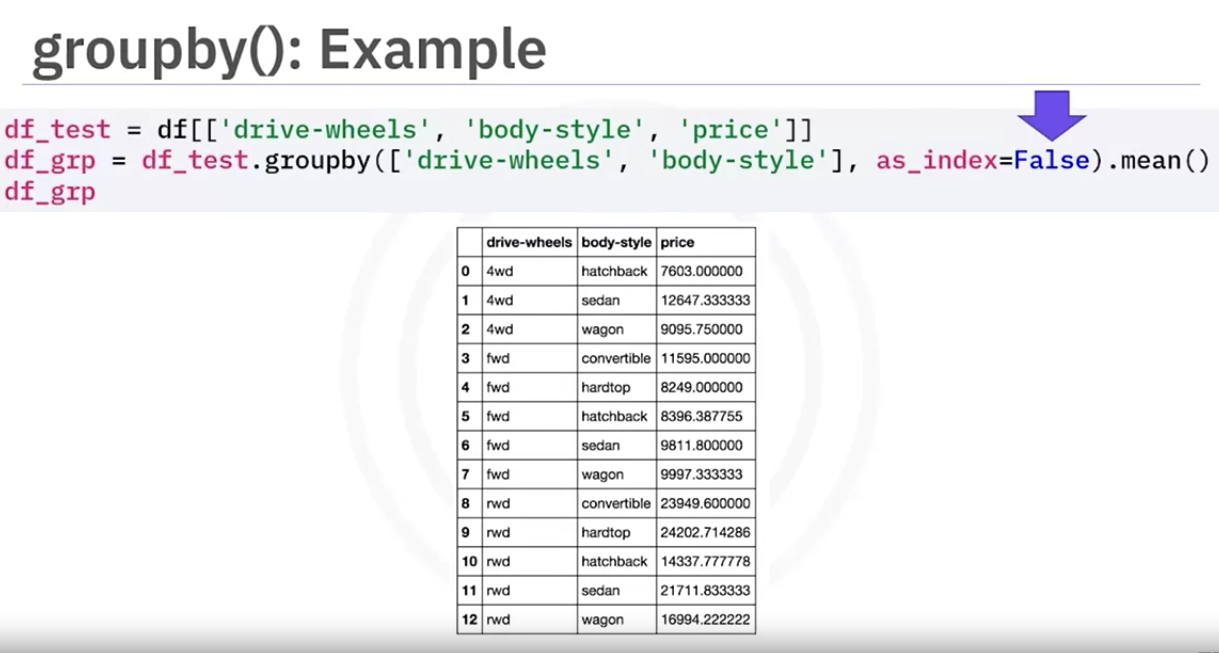
Can be used on categorical variables to group data into subsets according to different categories of the variable.

can group by single or multiple variables.

can apply aggregate functions on top of it.

df\_test = df.groupby([col1, col2,...], as\_index = False).mean()

Sometimes



The above data is hard to understand there is quite a lot of duplication inthe first column

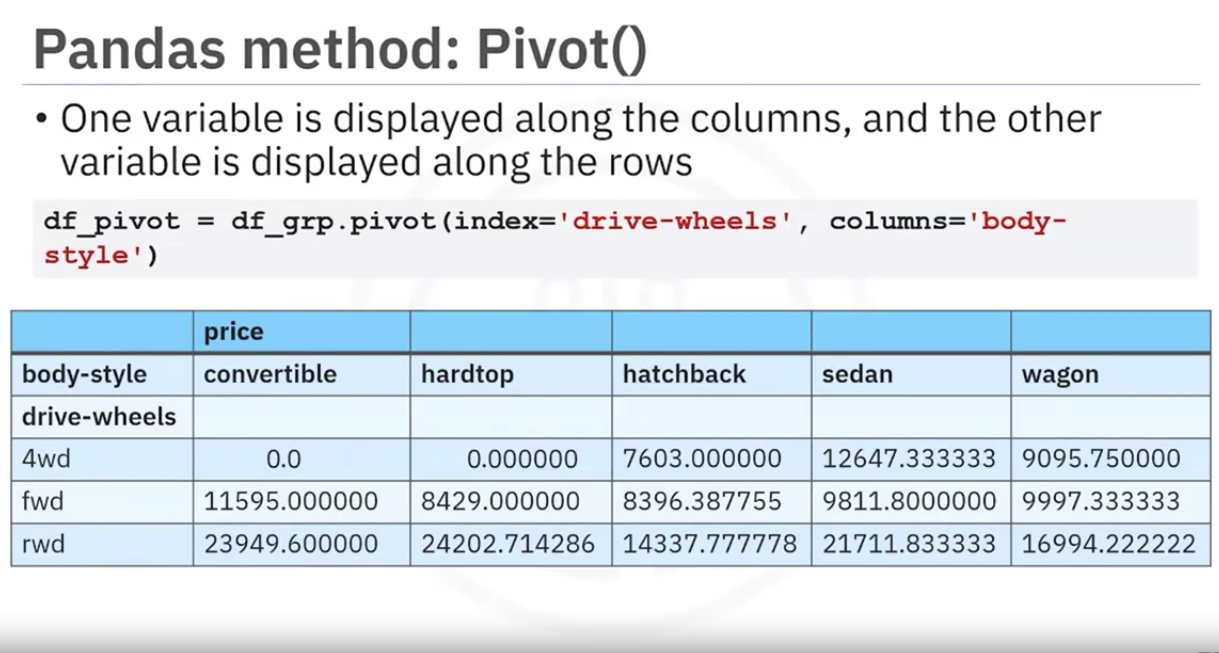
To better understand this, we can use a pivot table totransform the table.

Pandas pivot method

In the prevoious table,both drive whees and body style are displayed in columns bit with pivot we can represent one variable is along columns and the other variable along rows

df\_pivot = df.grp.pivot(index=”index\_name”, columns=”column\_name”)

This can be used when it seams like there is a crossjoin of sorts between two columns,



As we can see the drive wheel is in the index and body style has different categories in the differnt columns and we are displaying price for each combination.

One way to visualize the pivot table is by using a heatmap plot.

Heatmap Plot

Takes a ractangular gird data and applies color intencity based on the data value at the grid points

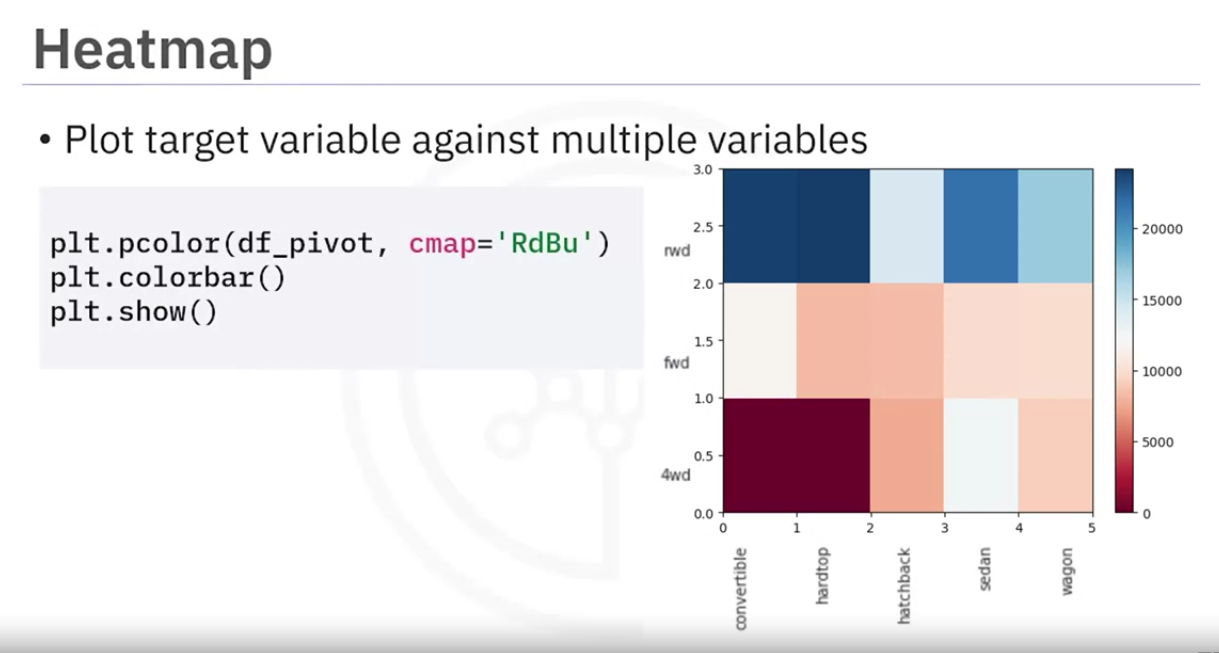
It is a great way to plot a target variable agains multiple variables and through this, get visual clues of relationship between these varuables and the target.

Weuse pyplot’s pcolor method

plt.pcolor(df\_pivot, cmp = ‘RdBu’)

plt.colorbar()

plt.show()



x axis is body style

yaxis is frove wheel

the average price is pplotte with vaying colors based on thier values according to the color bar.

Go to the reading1.

Note: After reading1 learn more about when to use the differnt graphs

Correlation

Measures to what extent differnt variables are interdependent.

Correlation vs causation chatgpt

needs 2 variables.

correation doesnt mean causation

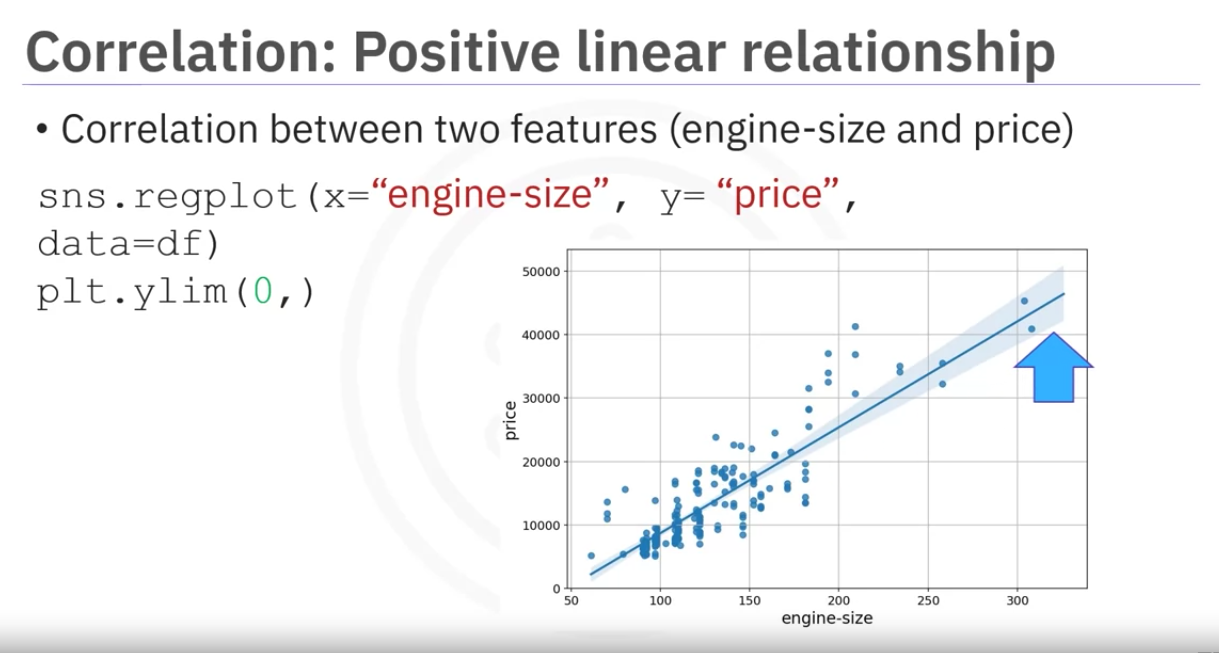
we use regression plot

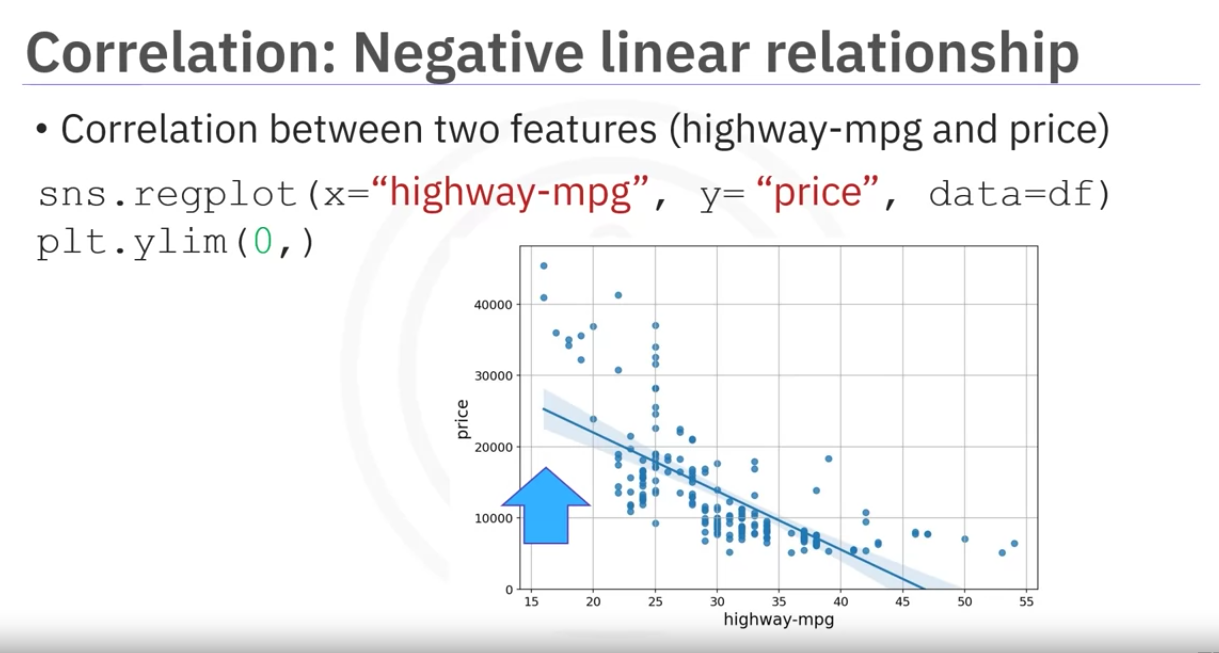
3 kinds of relationships

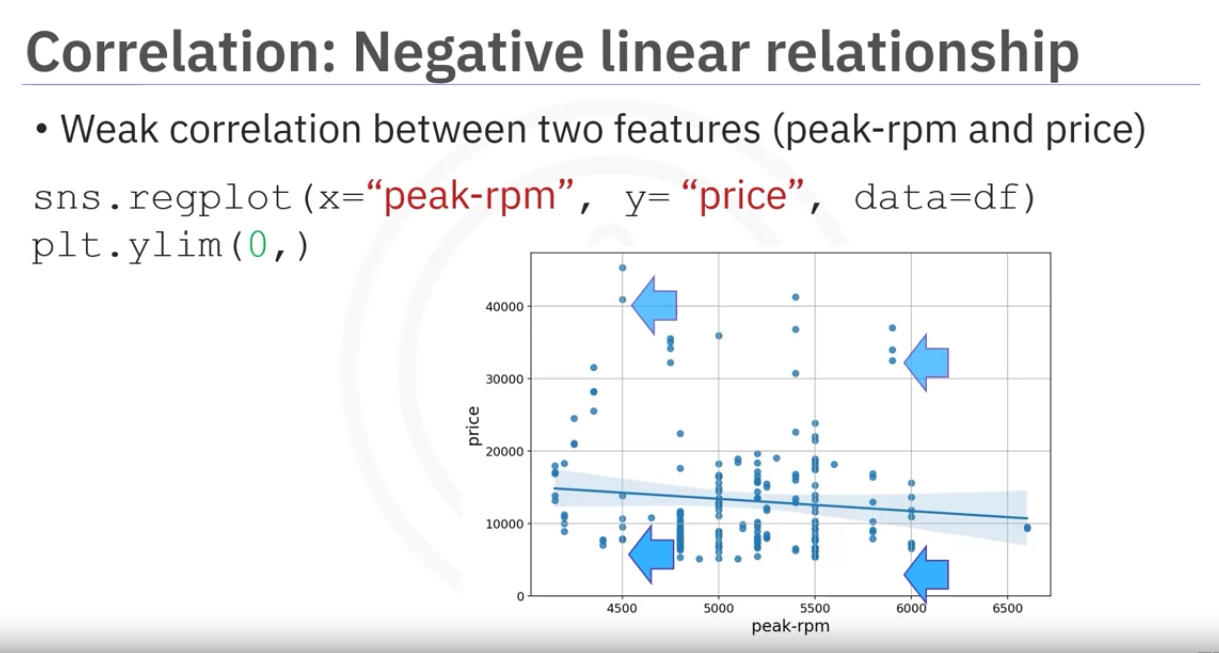
1 positive

2 negative

1. weak relationship







Correlation Statistics

There are different methods that allow us to calculate correlation between two variables

Pearson Correlation

Measures the strength of the correlation between two continuous numerical variables

Gives 2 values as result

correlation coefficient - between -1 to 1 . -1 is negative correlation. 1 is positive . 0 is no relationship

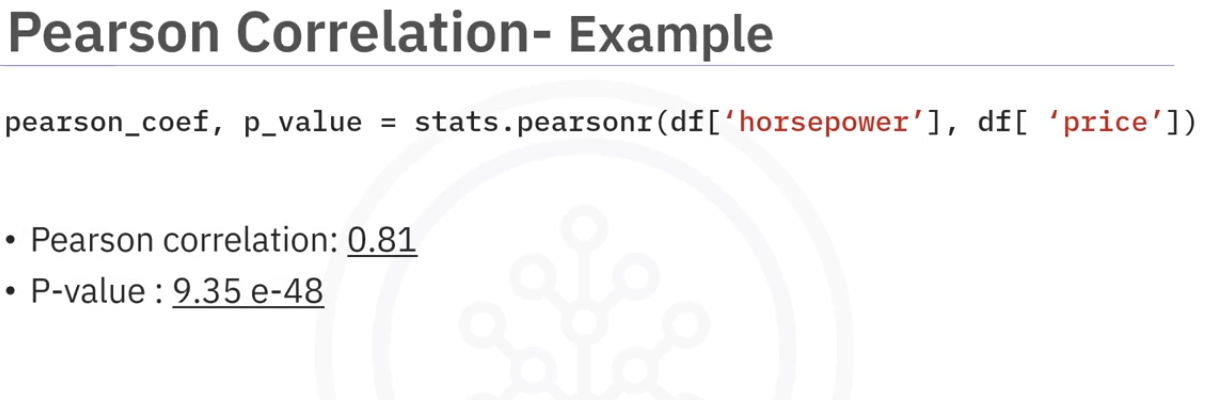
p-value - Tells how certain are we about the correlation calculated.

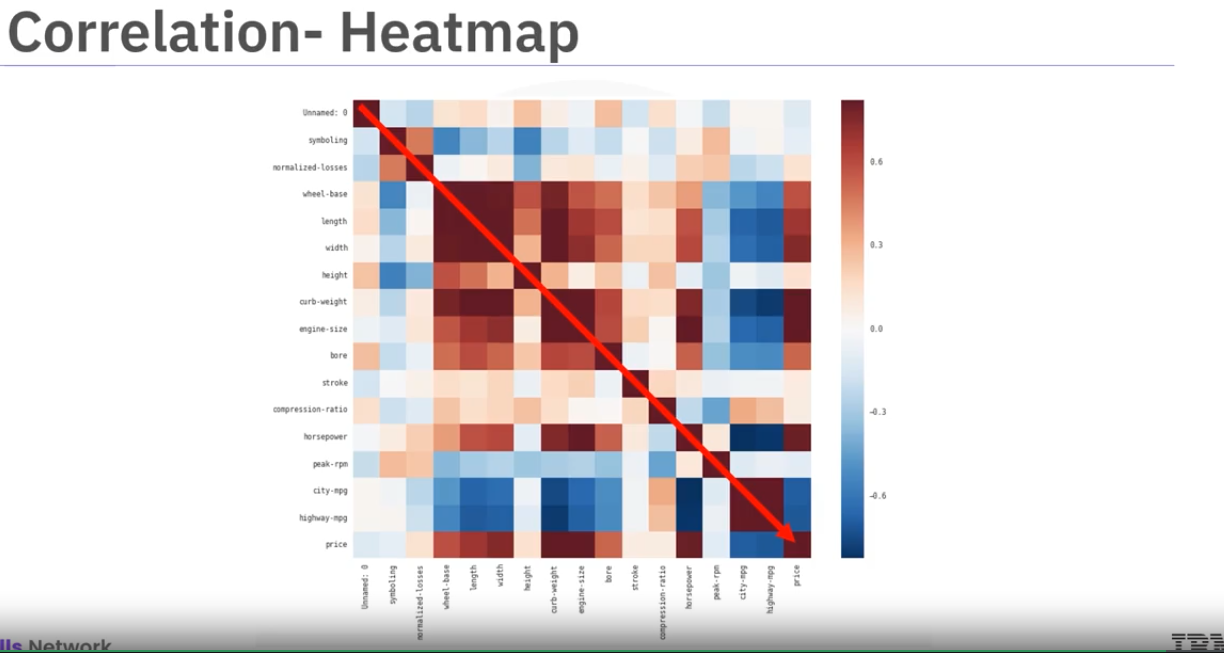
if less than 0.001, Strong certainity in the result

if less than 0.005 moderate

if less than 0.1 weak

if greater than 0.1 then no certainity,





# 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.