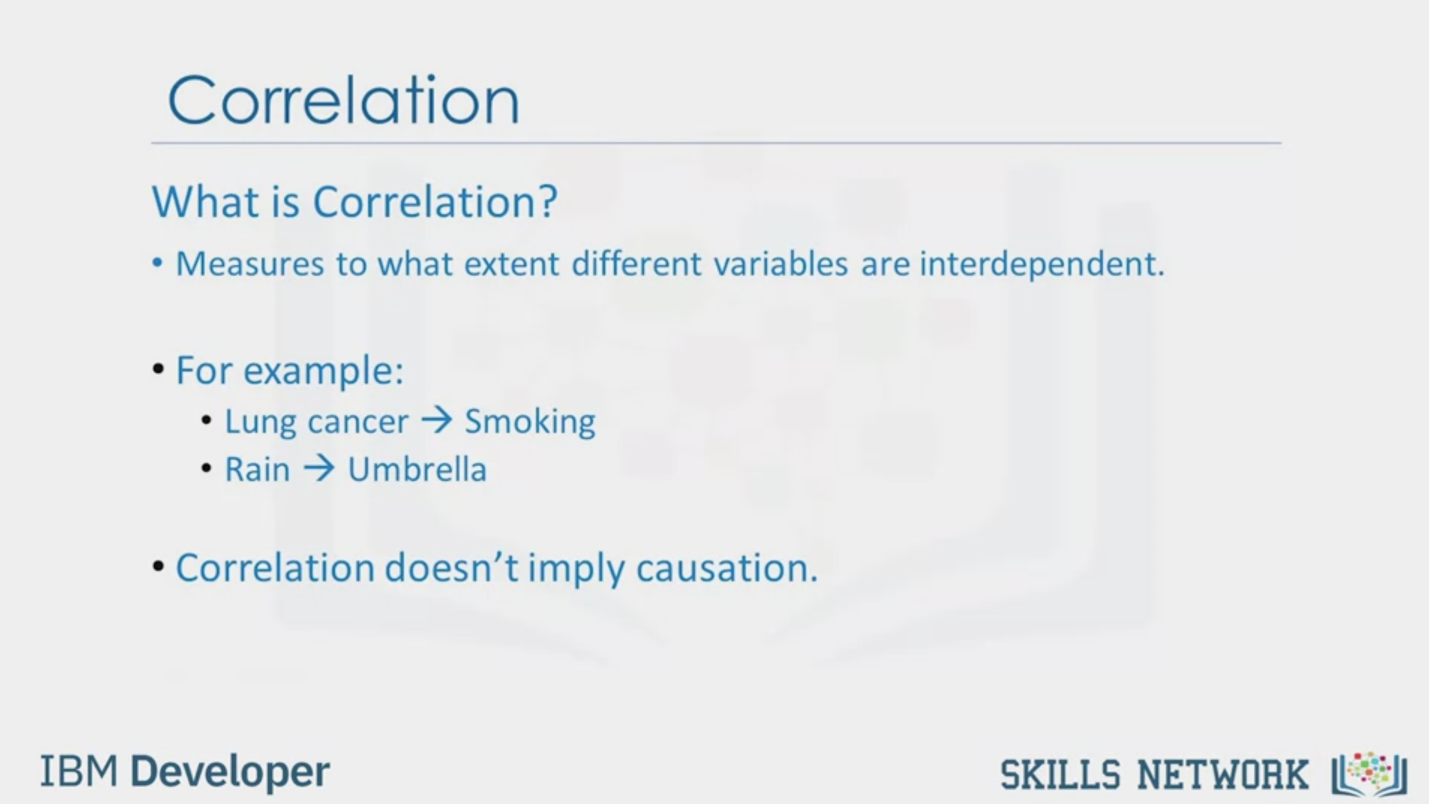


Correlation

IBM Developer

SKILLS NETWORK



Correlation

What is Correlation?

• Measures to what extent different variables are interdependent.

• For example:

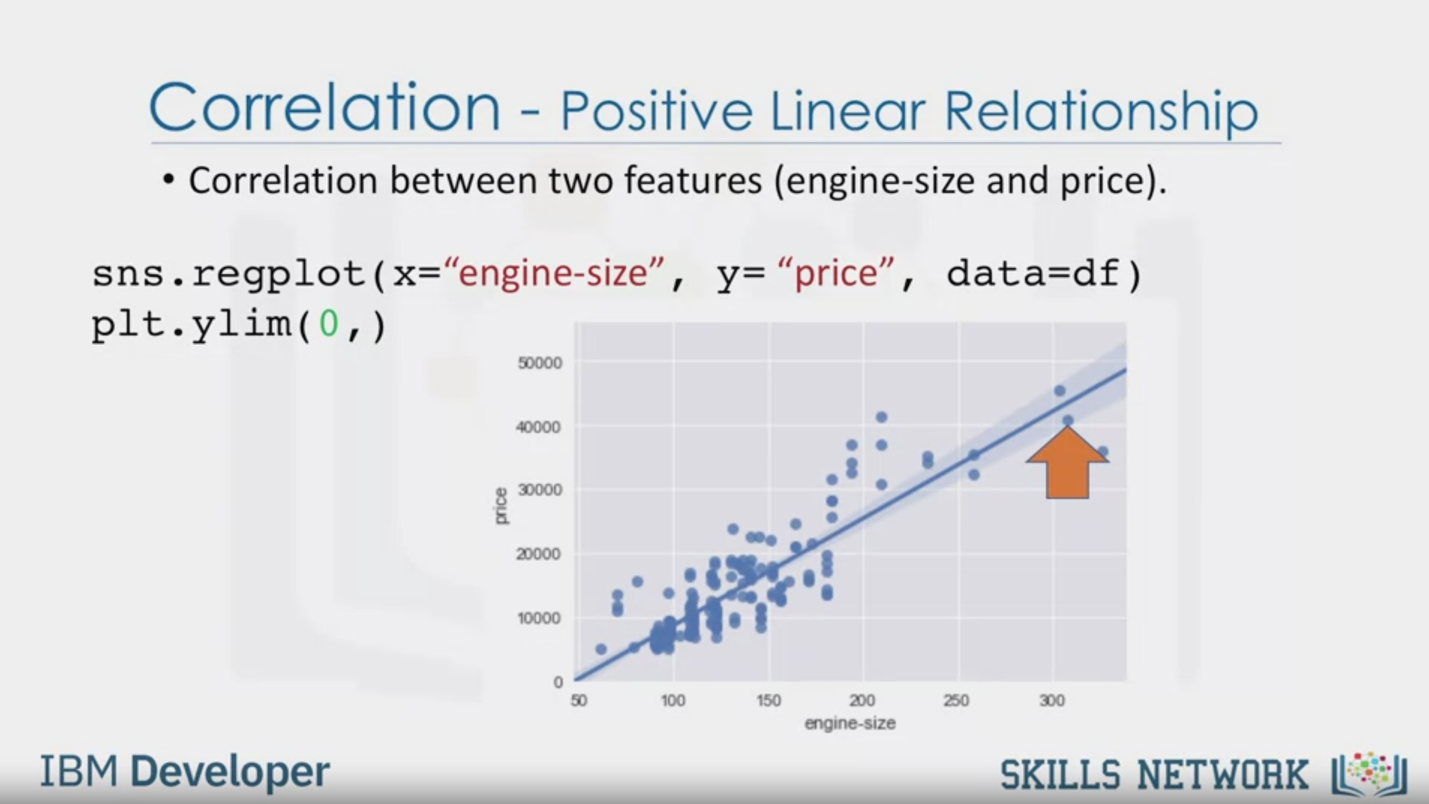
• Lung cancer > Smoking

• Rain > Umbrella

• Correlation doesn't imply causation.

IBM Developer

SKILLS NETWORK



Correlation - Positive Linear Relationship

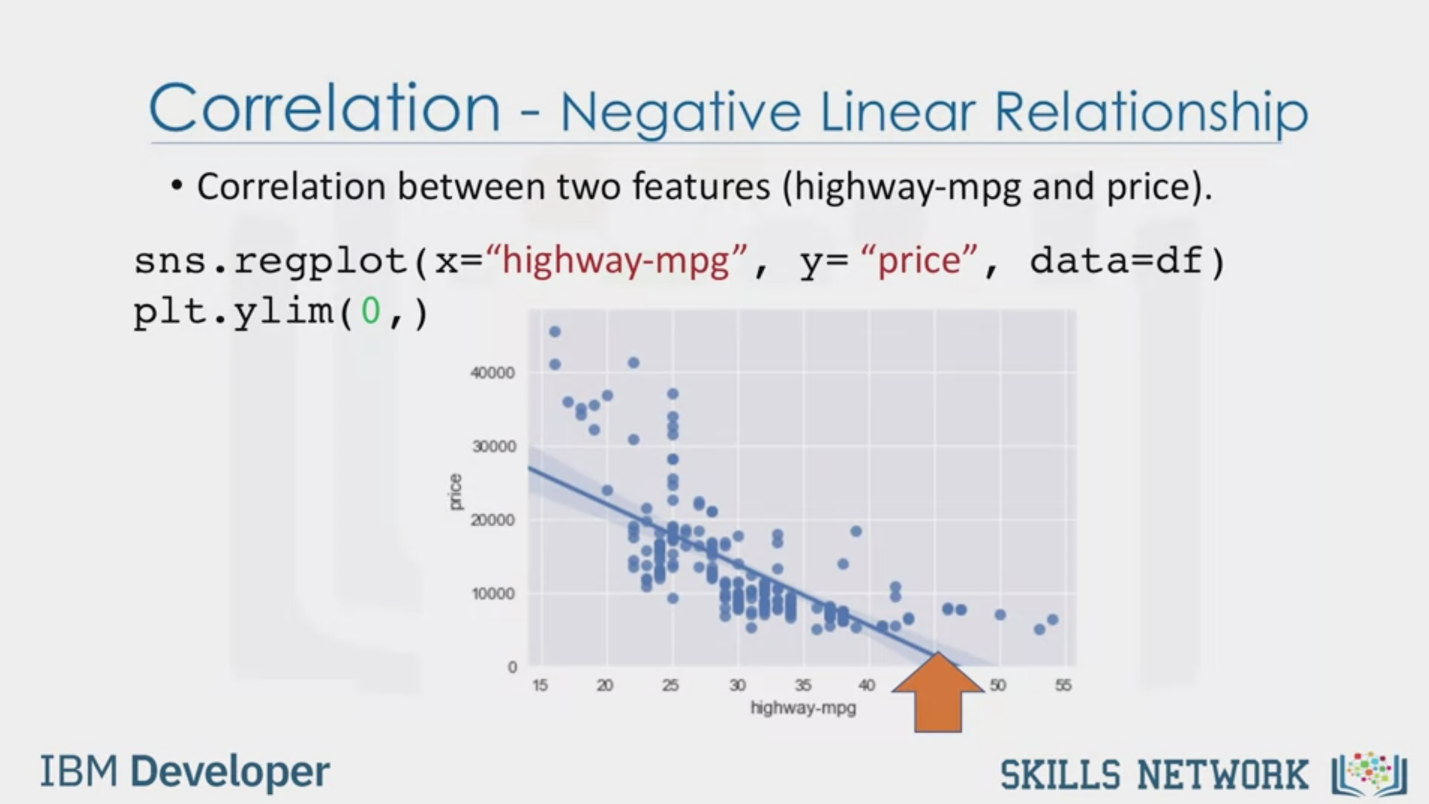
• Correlation between two features (engine-size and price).

sns.regplot(x="engine-size", y= "price", data=df)

plt.ylim(0,)

IBM Developer

SKILLS NETWORK



Correlation - Negative Linear Relationship

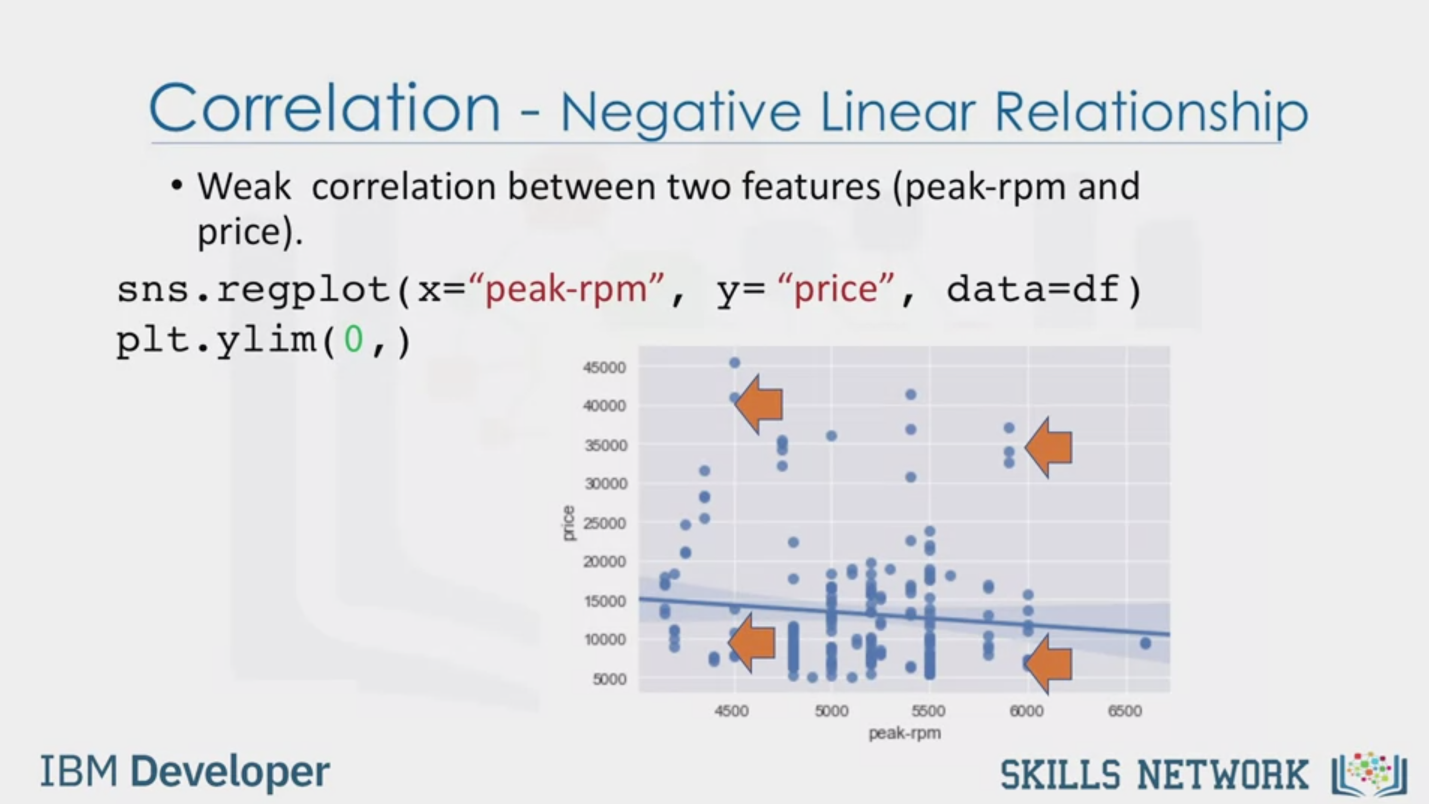
• Correlation between two features (highway-mpg and price).

sns.regplot(x="highway-mpg", y= "price", data=df)

plt.ylim(0,)

IBM Developer

SKILLS NETWORK



Correlation - Negative Linear Relationship

• Weak correlation between two features (peak-rpm and

price).

sns.regplot(x="peak-rpm", y= "price", data=df)

plt.ylim(0,)

IBM Developer

SKILLS NETWORK



SKILLS NETWORK

IBM Developer

In this video, we'll talk about the correlation between different variables.

Correlation is a statistical metric for

measuring to what extent different variables are interdependent.

In other words, when we look at two variables over time,

if one variable changes how does this affect change in the other variable?

For example, smoking is known to be correlated to lung cancer

since you have a higher chance of getting lung cancer if you smoke.

In another example, there is a correlation between umbrella and

rain variables where more precipitation means more people use umbrellas.

Also, if it doesn't rain people would not carry umbrellas.

Therefore, we can say that umbrellas and rain are

interdependent and by definition they are correlated.

It is important to know that correlation doesn't imply causation.

In fact, we can say that umbrella and rain are correlated but we would not have

enough information to say whether the umbrella

caused the rain or the rain caused the umbrella.

In data science we usually deal more with correlation.

Let's look at the correlation between engine size and price.

This time we'll visualize these two variables using

a scatter plot and an added linear line called a regression line,

which indicates the relationship between the two.

The main goal of this plot is to see whether the engine size has any impact on the price.

In this example, you can see that the straight line through the data points is very

steep which shows that there's a positive linear relationship between the two variables.

With increase in values of engine size,

values of price go up as well and the slope of the line is positive.

So there is a positive correlation between engine size and price.

We can use seaborn.regplot to create the scatter plot.

As another example, now let's look at the relationship between

highway miles per gallon to see its impact on the car price.

As we can see in this plot,

when highway miles per gallon value goes up the value price goes down.

Therefore there is a negative linear relationship

between highway miles per gallon and price.

Although this relationship is negative the slope of the line is steep

which means that the highway miles per gallon is still a good predictor of price.

These two variables are said to have a negative correlation.

Finally, we have an example of a weak correlation.

For example, both low peak RPM and high values of peak RPM have low and high prices.

Therefore, we cannot use RPM to predict the values.