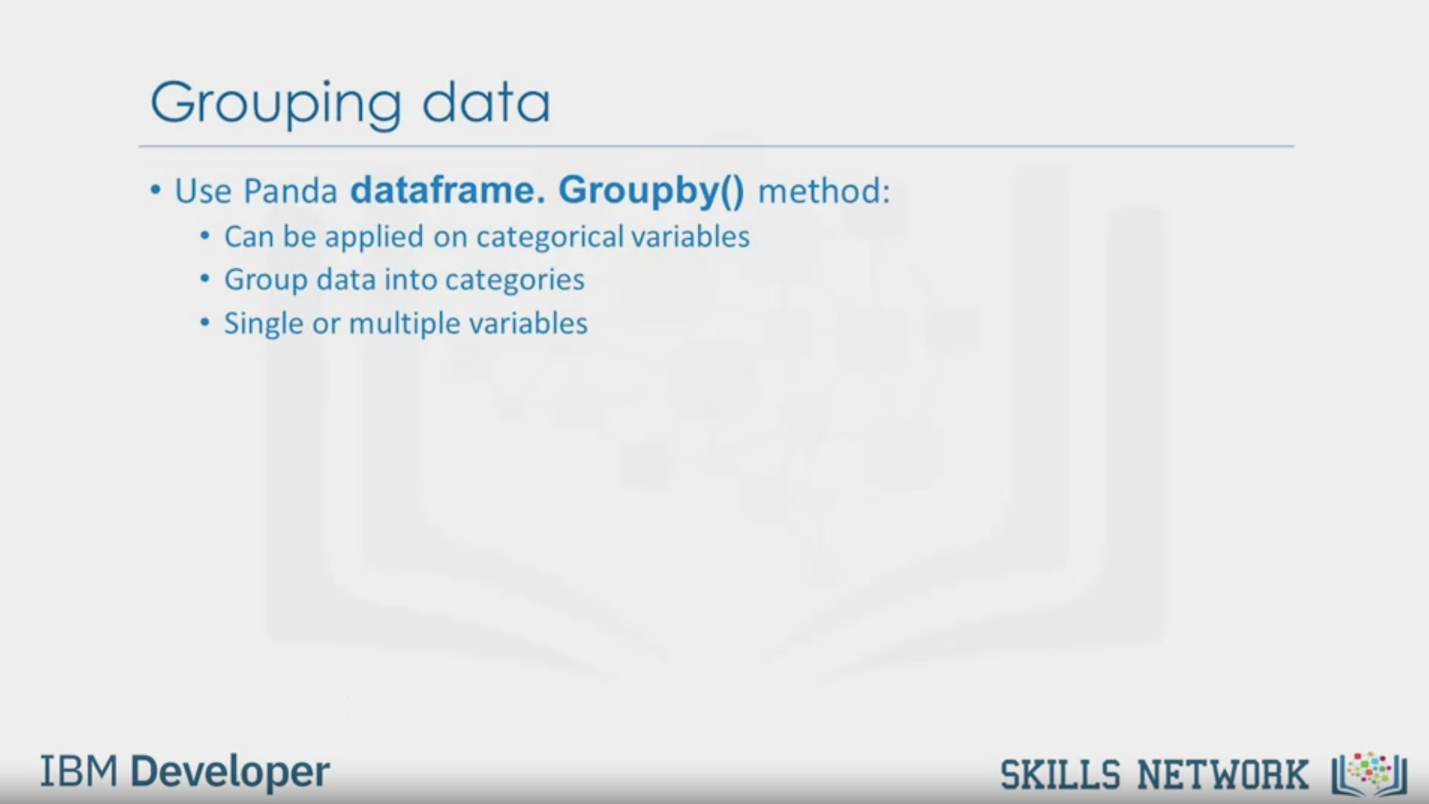


GroupBy in Python

IBM Developer

SKILLS NETWORK



Grouping data

• Use Panda dataframe. Groupby() method:

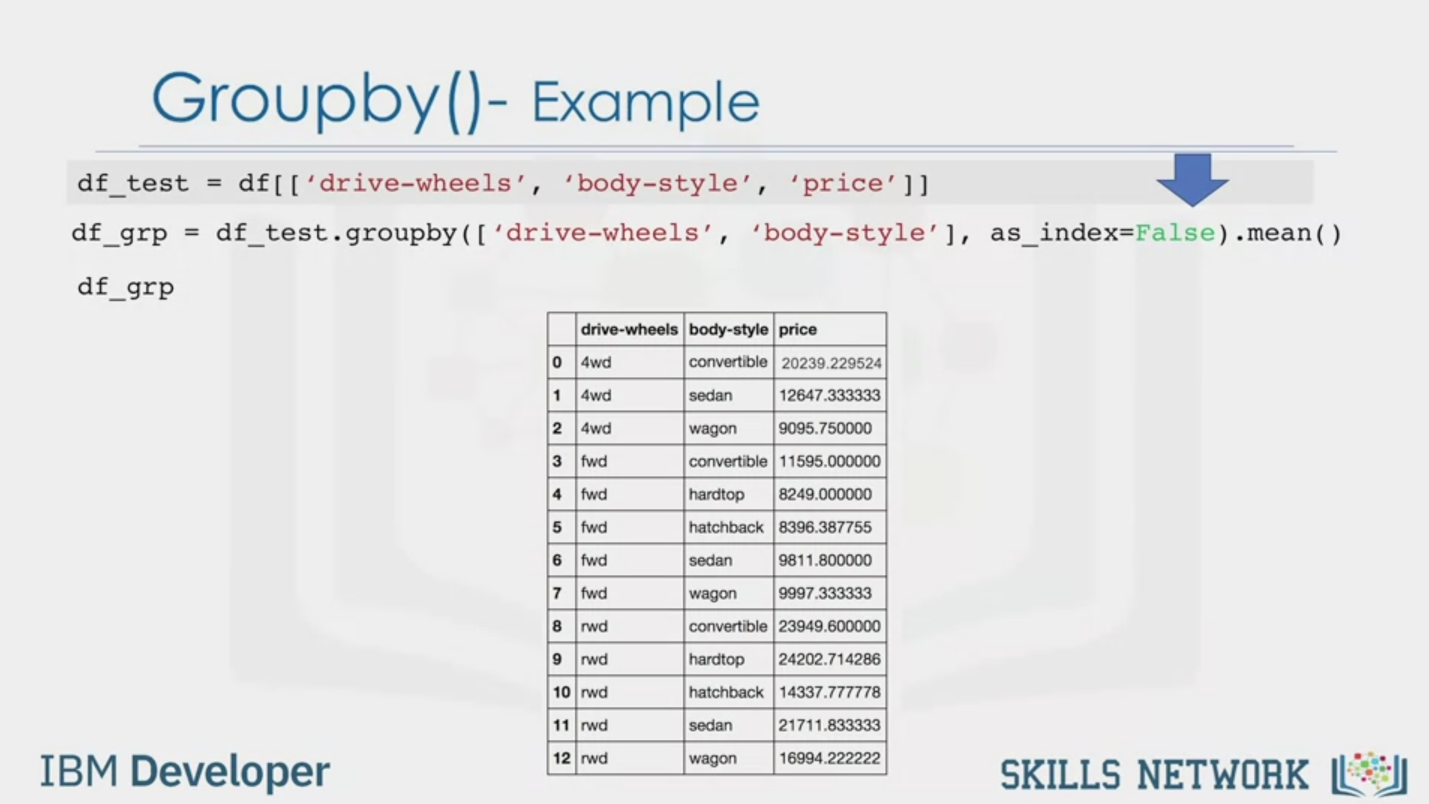
• Can be applied on categorical variables

• Group data into categories

• Single or multiple variables

IBM Developer

SKILLS NETWORK



Groupby ()- Example

df\_test = df[['drive-wheels', 'body-style', 'price']]

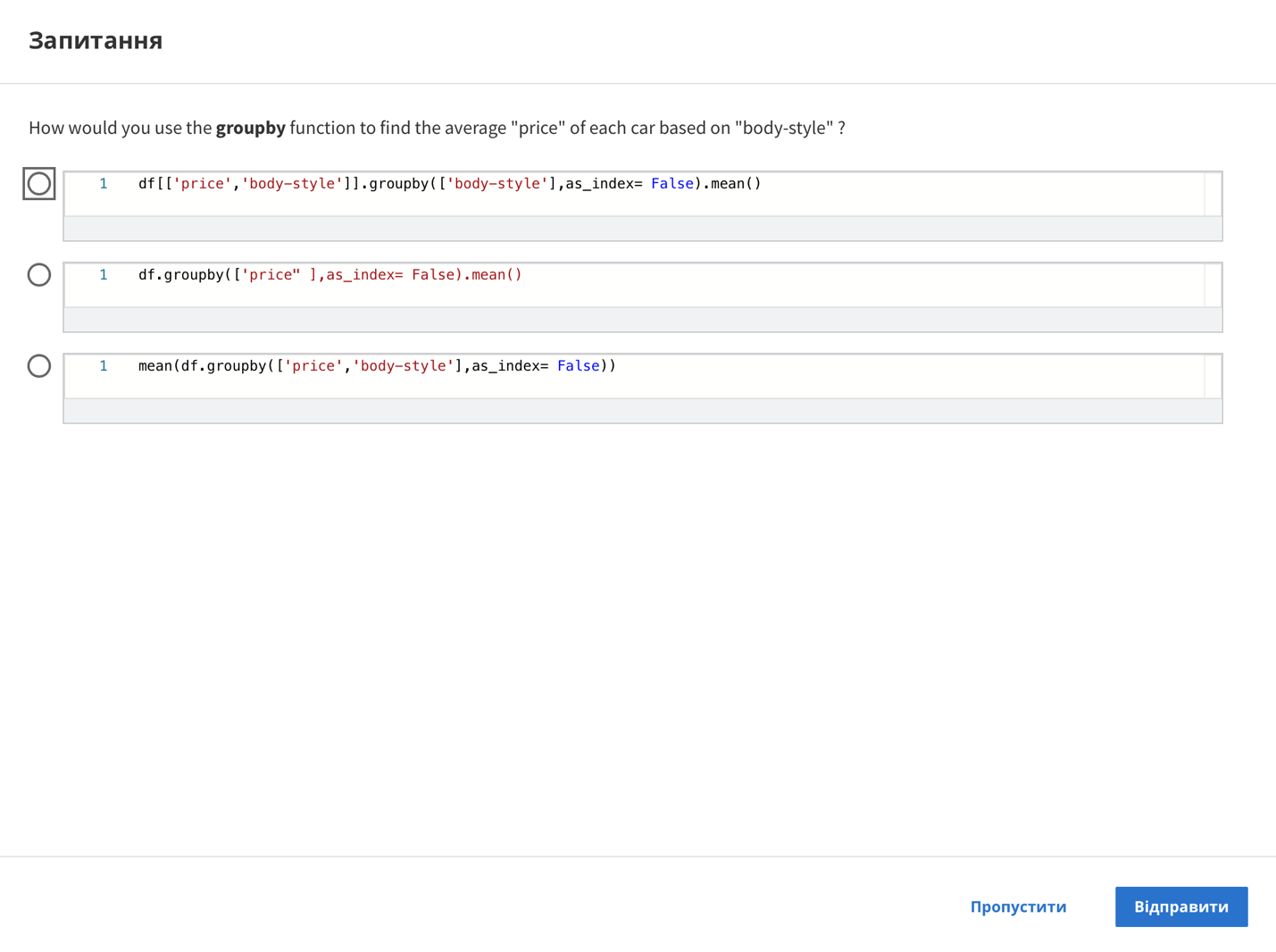
df\_grp = df\_test.groupby ([ 'drive-wheels',

'body-style'], as\_index=False).mean( )

df\_grp

IBM Developer

SKILLS NETWORK



Question

How would you use the groupby function to find the average "price" of each car based on "body-style"?

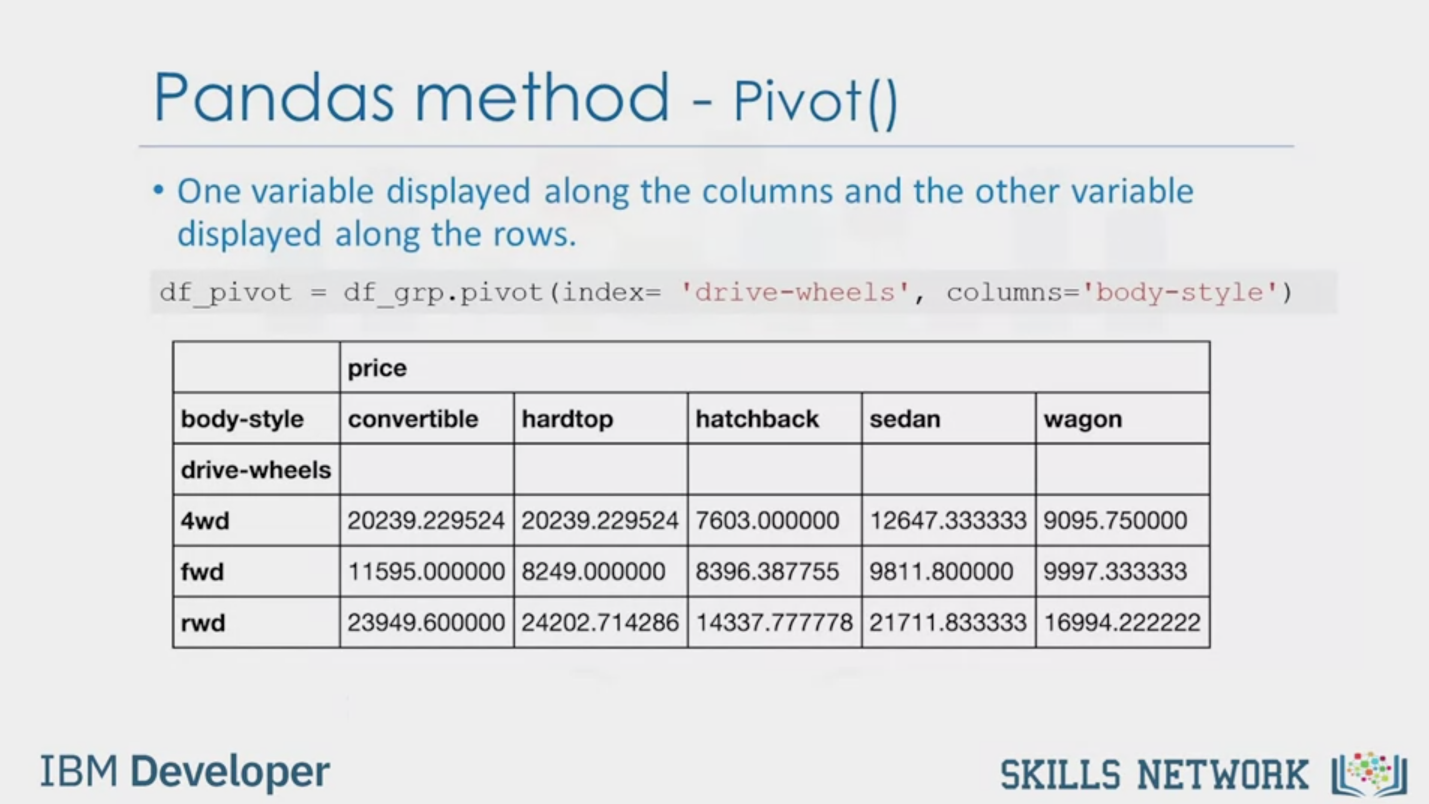
df[['price', 'body-style']].groupby(['body-style'] ,as\_index= False).mean()

df.groupby(['price"],as\_index= False).mean()

mean(df.groupby(['price', 'body-style'] ,as\_index= False))

Skip

Send



Pandas method - Pivot()

• One variable displayed along the columns and the other variable

displayed along the rows.

df\_pivot = df\_grp.pivot (index= ‘drive-wheels’, columns='body-style')

price

convertible

body-style

drive-wheels

4wd

fwd

rwd

hardtop

hatchback

sedan

wagon

20239.229524 20239.229524 7603.000000

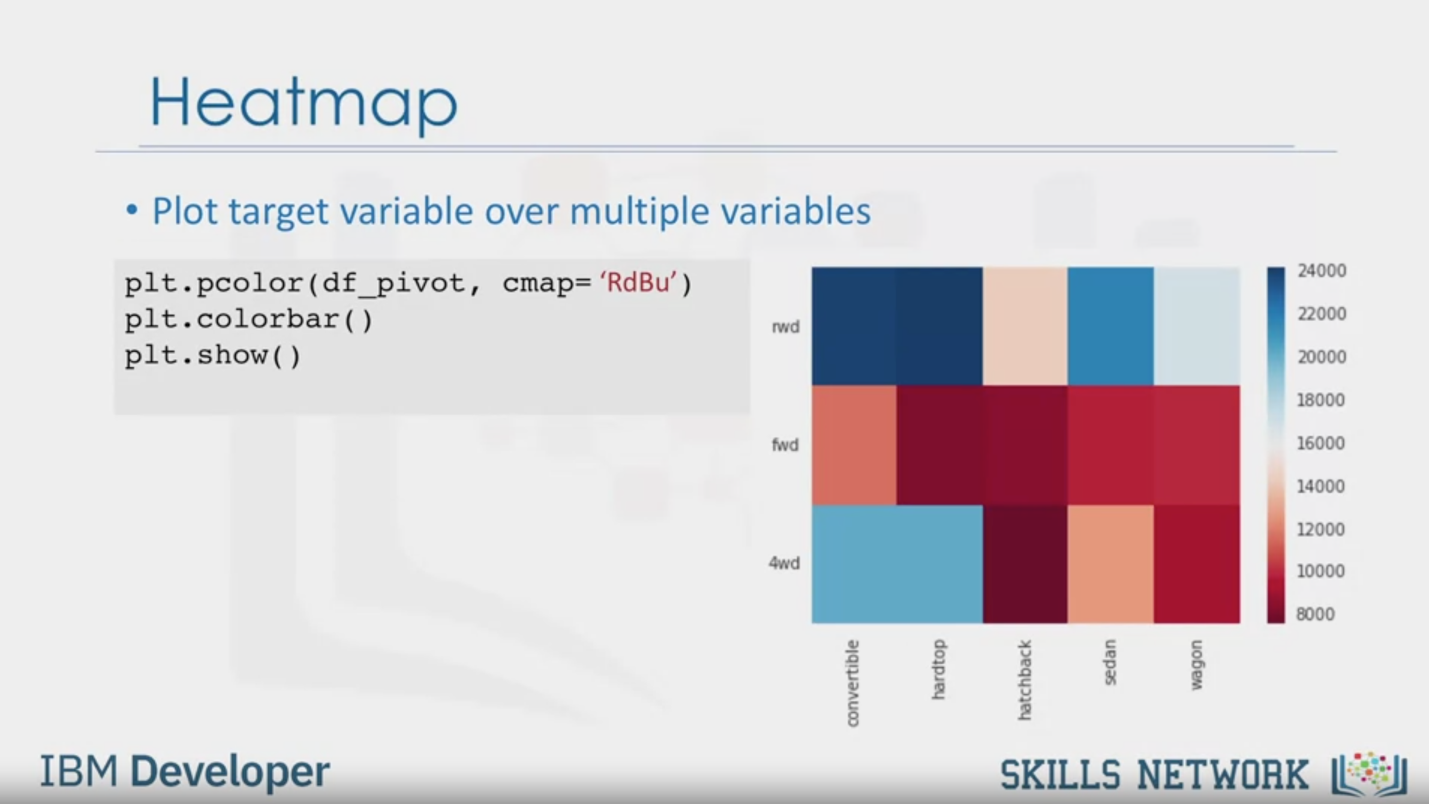
12647.333333 9095.750000

11595.000000 8249.000000 8396.387755 9811.800000 9997.333333

23949.600000 24202.714286 14337.777778 21711.833333 16994.222222

IBM Developer

SKILLS NETWORK



Heatmap

• Plot target variable over multiple variables

plt.pcolor(df\_pivot, cmap= 'RdBu')

plt.colorbar()

plt.show()

IBM Developer

SKILLS NETWORK



SKILLS NETWORK

IBM Developer

In this video, we'll cover the basics of

grouping and how this can help to transform our dataset.

Assume you want to know, is there

any relationship between the different types of drive system,

forward, rear, and four-wheel drive,

and the price of the vehicles?

If so, which type of drive system adds the most value to a vehicle?

It would be nice if we could group all the data by the different types of

drive wheels and compare the results of these different drive wheels against each other.

In Pandas, this can be done using the group by method.

The group by method is used on categorical variables,

groups the data into subsets according to the different categories of that variable.

You can group by a single variable or you can group by

multiple variables by passing in multiple variable names.

As an example, let's say we are interested in finding the average price of vehicles and

observe how they differ between different types of

body styles and drive wheels variables.

To do this, we first pick out the three data columns we are interested in,

which is done in the first line of code.

We then group the reduced data according to

drive wheels and body style in the second line.

Since we are interested in knowing how the average price differs across the board,

we can take the mean of each group and

append it this bit at the very end of the line too.

The data is now grouped into subcategories and

only the average price of each subcategory is shown.

We can see that according to our data,

rear wheel drive convertibles and rear wheel drive hard hardtops have

the highest value while four wheel drive hatchbacks have the lowest value.

A table of this form isn't the easiest to read and also not very easy to visualize.

To make it easier to understand,

we can transform this table to a pivot table by using the pivot method.

In the previous table,

both drive wheels and body style were listening columns.

A pivot table has one variable displayed along

the columns and the other variable displayed along the rows.

Just with one line of code and by using the Panda's pivot method,

we can pivot the body style variable so it is displayed

along the columns and the drive wheels will be displayed along the rows.

The price data now becomes a rectangular grid,

which is easier to visualize.

This is similar to what is usually done in Excel spreadsheets.

Another way to represent the pivot table is using a heat map plot.

Heat map takes a rectangular grid of data and assigns

a color intensity based on the data value at the grid points.

It is a great way to plot the target variable over multiple variables and

through this get visual clues with

the relationship between these variables and the target.

In this example, we use pyplot's p color method to

plot heat map and convert the previous pivot table into a graphical form.

We specify the red-blue color scheme.

In the output plot,

each type of body style is numbered along

the x-axis and each type of drive wheels is numbered along the y-axis.

The average prices are plotted with varying colors based on their values.

According to the color bar,

we see that the top section of the heat map

seems to have higher prices than the bottom section.