**Pandas DataFrames Concatenation and Merging**

[00:00:00.00] [MUSIC PLAYING]

[00:00:08.83] RYAN AHMED: Hello, everyone, and welcome to this lesson on Pandas DataFrame, concatenation, and merging. In the real world, data comes from many sources and often in many individual files. Sometimes we wanted to combine these files into a single Pandas DataFrame to analyze the data further or maybe train a machine learning model as an example. Pandas offers many methods to combine DataFrames, including merging and concatenation.

[00:00:39.17] Here are the key learning objectives of this lesson. perform DataFrame concatenation using Pandas concat function. Perform DataFrame merging based on a given column using Pandas merge option. Reset Pandas DataFrame index using the reset index method. So let's head over to our Jupyter Notebook and get started.

[00:01:02.42] [MUSIC PLAYING]

[00:01:09.79] All right so right now we are in the Pandas DataFrame, concatenation, and merging Jupyter notebook. So let's go ahead and get started. First I wanted to start by covering Pandas DataFrame concatenation. So Pandas DataFrame concatenation works by combining two or more DataFrames along a specific axis to produce a new one.

[00:01:33.25] For example, let's assume that I have these two Pandas DataFrames, investor\_1\_df and investor\_2\_df. And I want it to perform concatenation or combining these two Pandas DataFrames together to generate a combined DataFrame. And here I'm going to call it investor\_df\_combined for example.

[00:01:55.69] To do that operation in Pandas is actually pretty straightforward. All you need to do is to say pd.concat, which stands for the concatenation method. You open parentheses. You open square brackets. And then you specify the names of the DataFrames that you need to combine or concatenate together. And you're going to put that in a combined Pandas DataFrame afterwards. And that will be the form of my combined DataFrame. So simply you're going to put investor\_2\_df underneath investor\_1\_df.

[00:02:30.27] And I'm going to show you as well how we can go ahead and perform a reset to the index. So here you will notice that the index starts with 0, 1, 2, 3, up until 8. And then the next one, which is investor\_2\_df, came here afterwards below it. And the index here is 0, 1, 2, 3, and 4. Obviously when we combine them, we would like to reset the index kind of to start from 0 up until the maximum number at the end here at the last row. And I'm going to show you how we can do that in code coming up next.

[00:03:02.97] All right. So the next operation that I wanted to show you is Pandas DataFrame merging. Let's assume that here I have my investor\_df combined. This is simply the Pandas DataFrame that I created in the previous step. And let's assume that I was able to get additional information about all my clients, such as their salary information, as an example, and maybe their job title, for instance.

[00:03:29.47] And now I wanted to add that new information or merge that new information to my investor\_df\_combined. To do the merging operation, I need to specify which column I'm going to merge my data on. For example, if you check out the investor\_df\_combined here, you will find that simply here I have the investor ID is a unique ID for each individual investor. And the same ID as well is included here in the investor\_new\_df. So in order to combine these two together or merge these two together, I can merge the two Pandas DataFrames based on the investor ID column, because it's a common column between the two Pandas DataFrames.

[00:04:18.07] So to do that in code, all you need to do is to say pd.merge. That is going to be the merge method. You specify the first Pandas DataFrame. That is the investor\_df\_combined. You specify the second DataFrame, which is going to be investor\_new\_df. That's the new information, which is a salary and job title that I was able to obtain.

[00:04:39.28] And you need to specify the on, which column you wanted to merge these two data frames on. And that is going to be the investor ID. And once you do that, you're going to put that data in a new Pandas DataFrame, merged Pandas DataFrame. And here I'm going to call it investor\_df.

[00:04:57.25] So let's go ahead and show you how we can do that in code. First I want you to go ahead and restart the kernel. So if you say kernel restart and clear output, that is going to clear all the outputs for us. And then I wanted to import Pandas. We learned how to do that many times before. So I'm going to say import Pandas as PD.

[00:05:16.48] Next I wanted to read two Excel sheets. The first one is going to be investorsgroup1.csv. And please note that the investors group 1 and 2, both of them here are included in the same location as my Jupyter notebook. So if I say pd.read\_csv investorgroup1.csv, I'm going to put my data in a Pandas DataFrame. I'm going to call it investor\_1\_df.

[00:05:46.48] If you press Shift Enter, here we go. If you zoom out a little bit, you will see here I have my first Pandas DataFrame. So here I have the investor ID. And this is a unique ID for each of the investor. Here I have the first name. I have the last name, the age, portfolio size, years with investment firm. I also have the risk tolerance and the goal as well.

[00:06:09.87] Next I'm going to go ahead and read another Excel sheet titled investors\_group2.csv. I'm going to say pd.read CSV. I'm going to put that data in investor\_2\_df. And you will notice that this is simply another group or batch of customers that I have in another Excel sheet. Again, each of these customers have a unique ID, B1, 2, 3, 4, and 5. These have different names and the exact same column names as well.

[00:06:41.63] And next what I wanted to do is I wanted to simply combine these two together. I wanted to do Pandas DataFrame concatenation. I want to take investor\_2\_df. I want to put it underneath of investor\_1\_df. To do that, I'm going to say pd.concat. You open parentheses. You open square brackets. And then you list these two Pandas DataFrame here within square brackets. And you're going to that data in investor\_df\_combined.

[00:07:10.63] So if you press Shift Enter, here we go. What you notice is now I was able to put my second Pandas DataFrame here underneath the first one. And of course, you'll notice as well on the left hand side that the index here for the original raw data frames are still kept intact. For example, here I started with 0, 1, 2 up until 8. And then I restarted again at 0 up until 4.

[00:07:34.03] To reset the index, all that you need to do is to grab the name of the Pandas DataFrame, investor\_df\_combined. And if you say .resetindex and you say drop equals to true, if you press Shift and Enter, there we go. What you notice is now I was able to reset simply my index. And that was pretty straightforward. So I started with index 0 up until index 13.

[00:08:01.16] And next what I wanted to do is I wanted to show you the merge operation. Let's assume that we acquired new information about my investors, such as their salaries and maybe the job title as well. So if I say pd.read CSV and if I grab my investorsnewinformation.csv, this is simply here another CSV file that I've included as well. It's called Investor New Information.

[00:08:24.98] I'm going to read that data, and then I'm going to put it in a new Pandas DataFrame called investor\_new\_df. This is, again, an additional new information about all my clients, all my combined clients. You will notice the investor ID here as well is unique for each one of them.

[00:08:42.66] And now we can simply do Pandas DataFrame merging. I want to merge these two Pandas DataFrames based on the investor ID column. To do that, you say pd.merge. You grab the investor\_df\_combined, the first one. You add the new information to it, and you specify which column you want to merge on.

[00:09:03.53] So on is going to be equals to investor ID. And you're going to put that in investor\_df. If you press Shift Enter, here we go. Now you have been able to come up with this kind of combined and merged Pandas DataFrame containing all my clients here along with the additional information that has been captured. So here I have the salary and the job title as well.

[00:09:24.86] All right, so that's it. That's all I have for this lesson. I hope you enjoyed it. In the next lesson, we can have our practice opportunity. Please go ahead give it a try. And then I'm going to show you a detailed video of the practice opportunity solution. I hope you enjoy this lesson, and see you in the next one.

[00:09:41.51] [MUSIC PLAYING]