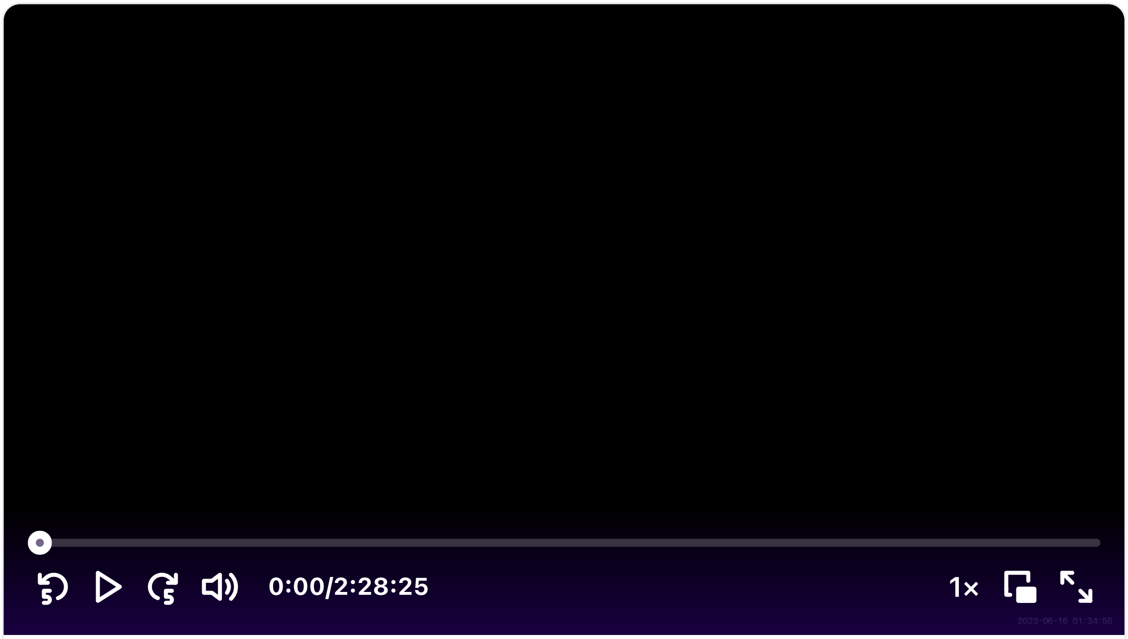
Day 11 Exploring Pandas



Select text to create highlights Learn more

Transcript

S1 Speaker 1 0:04

Thank you for recording the, the meeting and, and we're gonna, we are gonna see that part. Okay, so you can't see my screen. You can see my Python. You can see my Jupyter. Okay, so we went through Jupyter. So we, we know how to open Jupyter and I'm gonna start with my first row of code that is gonna be import pandas as pd. Then I'm gonna create my path file resources, Latin Street dot psp. So if I go here in and I go here to my first activity, I am in the folder section, you're gonna see that resources is in the same level as my Jupiter. So that's why I can put my, like this resources and back on the street dot. Okay. Again, I'm gonna repeat this again because see some of you are having some problems here. So basically when this is not in the same label, you have to go to that path in order to put it in the file path. Okay? One other option that you can do basically is that you can go to that path. Okay? I am going, you're going to do one, one quick example here. Go to that path, okay? For example, please copy the path from Explorer.

S2 **Speaker 2** ▷ 1:31

Sorry, you go. We missed the part where you copied something,

S1 **Speaker 1** ▷ 1:38

So, oh yes. Yeah, yeah.

S2 **Speaker 2** ▷ 1:40

So you are sharing only the Jupyter lab?

S1 **Speaker 1** ▷ 1:44

Yeah. Let me do it right now. Yeah. Can you see my folder?

S2 **Speaker 2** ▷ 1:50

Yes, we do.

S1 **Speaker 1** ▷ 1:51

Yeah. Okay. So basically you can go here for example here, and then you just click this pad, okay? And this is the string that you're going to use. Of course the string that you're going to use here. Basically you have to fix it, okay? Because you write it differently in pipe. You can see my Jupiter now? Yep. Yes. Perfect. So yes, and you use copy and paste. Okay. But I mean one of the easiest way to do it is just to make sure that you have whatever you are going to read in the same place where you are actually reading your, I mean that folder in the same place where you're actually reading your Jupyter. Or if you want, you can have that file this bottom street, E S B here, outside the resources and you just put the file path like this. Okay? Is that clear? So top, yeah, perfect. Okay, in that case then we go to the part with where we can actually read that file with read e, sb, and PDF with pandas file. And then I can print it original DF head, who can remind me what is head for

S3 **Speaker 3** ▷ 3:25

Read. The first five. First five rows.

S1 **Speaker 1** ▷ 3:28

Rows. Thank you very much. Yes, good. The five first rows. Okay, well now I have the five, five first rows and I can see my data frame, a street name id, street name, a street, full name and the column. So now I can do, I can start doing things right. I am going to work with I, I lock and lock, but with lock, remember I can't get my index indexes as an integer. So that's why I am going to define a new variable. If it's equal to original data frame dot. And I'm going to use the function set index. What column I am going to use to set my index? The street name. Why the street name? Well this is the exercise, okay? So when you print again the head, the street name is your index. Now it's not in the level of your header, but it's like the index.

S1 **Speaker 1** ▷ 4:24

So this means that instead of private street from a street full name becomes the row zero, it is gonna become the row private street. This is becoming the row fourth, the road 11, okay? But the index are there. So when using I lock, I still can access these three indexes. Ok? So now that I have my index, set it up. Remember this is similar from the pipe tables that we did in Excel. When we put this as my row column as my row, and I can actually populate information about that. So now I'm going to use lock and I lock, okay, I'm going to create a new variable. I is gonna recall adding Addington name, okay? What I want to bring here in Addington, name the frame block, my brackets notation. And what is this first? It's a row or it a column. Row. Row, exactly. Perfect. So my notation is row column. So lock, row, adding, accessing through an string. And column three, full name. So using lock, adding on name, I am going to print. Okay, copy is, that's clear. Question so far.

S3 **Speaker 3** ▷ 6:05

I have a question. I I missed the part. When you say that if once the street name, if it becomes a column, you mentioned that the index, what is the index? Zero then there,

S1 **Speaker 1** ▷ 6:16

No, the street name. We set street name as an index.

S3 **Speaker 3** ▷ 6:22

So street name is index zero and then private street is one and fourth is two and

S1 **Speaker 1** ▷ 6:28

11, three, no, here we're gonna start counting from private three.

S3 **Speaker 3** ▷ 6:32

Oh. So that's gonna be zero. And then one and two. 11 will be two. Oh, okay.

S1 **Speaker 1** ▷ 6:37

Got it. Exactly. Perfect. So I'm gonna do exactly the same thing, but now I am going to use a lock. So I think Tony is gonna be 0, 1, 2, 3. That's why I put first my row and the street full name is zero one. I am going to bring value. Okay, so then I can do something else. Okay? I can actually bring all the roles and the columns. And to do this, I am going to put, because I'm, I am bringing a list inside my, my lock. You see here, private, digital phone. I'm gonna bring my data frame, my data frame, this data frame, this is my data frame, okay? Lock, I mean don't lock. And here my same notation, guide, bracket, bracket, and my first list, what is my first list? Rose or columns?

S1 **Speaker 1** ▷ 7:54

Rose. Rose, perfect. My first list of rose and my after the coma, my second list of columns. So I am bringing everything that is in private street. Everything that I have in court, everything that I have in 11th and everything that I have in income. From which columns? From the street name id, street, full name, postal and postal community. So I am taking these and this code and I am bringing everything that has to do with private street, private street, 11th. 11th adding on fourth and child. So basically I am bringing everything. Okay, well not everything because here I, I'm just breathing head.

S1 **Speaker 1** ▷ 8:52

Okay, questions so far guys? So one thing, when you print, you're gonna get this account, okay? So with that filter that I apply, yes, with that filter, that that, that I apply with log I, I brought 329 rows and three columns. Okay? The second print that I'm gonna do is actually, I'm gonna do it exactly the same thing. But with I lock with I lock, I want to bring everything range with two point, that's my rows zero to 5, 0, 1, 2, 3, 5. And my columns are going to be 0, 2, 3, 0, 1, 2, 3. Okay? Not including three. Okay? Remember that Postal community, yes. Yeah. So I log seems faster because you use the indexes to bring everything. Log seems a little bit tedious because you have to type everything that you want to bring. But I am going to give you some useful examples when lock is really useful, okay? Next part is that actually when we put points in the lock in the rows, that means bring everything, all the rows that you have, okay? From these list of columns only this list of columns, straight, full name and postal community.

S1 **Speaker 1** ▷ 10:45

Okay? The other way around is gonna work the other way around. You can put two, two the semi column here and then the list of rows that you want to do. You can do actual, you can do with I lock and lock. You can do this kind of with conditionals. This is not using I lock and lock. This is the example that I can set up a variable al equal to accessing to an specific column. Unity community and this municipal community that is this one, it has to be equal to courage, okay? And then bring whatever I have in courage. So what I'm, what is, what this is going to be doing is going to check this first private history in municipal community has parish check falls Fault, fault, fault is gonna be bringing everything as as bullions. Okay? So the example was to actually use lock, I mean lock and I lock, but in this case I'm gonna use lock, first lock. Then I access to the same d F port community and equal equal prayer bill. Okay? Prayer bill basically is going to be my what? Rose or columns here guys, this condition is my rose or is my column rose Ropes. Ropes, exactly. I'm bringing all the cones as well here. So I am all the cones

S3 **Speaker 3** ▷ 12:46

Or the rows,

S1 **Speaker 1** ▷ 12:49

All the columns here.

S3 **Speaker 3** ▷ 12:50

Okay.

S1 **Speaker 1** ▷ 12:51

Right, because these two points in the right side, okay? And, and then, and we can actually put multiple conditions, how you're gonna put multiple conditions. Well basically you're gonna use this pipe symbol, okay? Pipe, pipe. So this pipe symbol, you just put your first lock, you put your first condition here, then put the pipe timble, put the second condition and then bring everything. Okay?

S3 **Speaker 3** ▷ 13:36

Why is not, oh yeah it is. And square bracket. Okay, got it, got it. Sorry.

S1 **Speaker 1** ▷ 13:40

Yeah, no problem here. The printing on, on, on Jupyter notebook is not that good. Like I mentioned, you're gonna have these kind of things going on. For example, here I can print more rows and they printed like three dots and that's why it looks like weird. So that's why when cleaning the information, we're gonna see how to fix that, exporting that into A C, S B. Okay? And then you can see everything More clear questions, guys?

S2 **Speaker 2** ▷ 14:23

Yes, I do. Yes. In line number seven, I guess that's the one, yeah. Yeah, this one. So you added colon in first where you will get all the rows and then you added comma where you will get the two columns, right? And then you added head again. So this won't, will it not confuse it because if you're adding colon, it will bring all the rules, right? And then head you are saying head, head will bring only five rules. So is it not contracting or how, how does it work?

S1 **Speaker 1** ▷ 15:05

Well, no, basically here because I wanted to show it like more clear in my printing, I am using my dark head. Okay. But because this instruction is to actually bring all the roles, it's two column only, I can actually start increment my head by 10, by a hundred by a thousand and so on. So I can be checking what is printing my head or my tail. Okay. And the the important part here is that I am indicating to bring all the rows for two specific columns. Okay? Yep, that is true. And that's the filter that I'm doing with lock.

S2 **Speaker 2** ▷ 15:57

Okay, got it. So if that head won't be there, then it would've printed all the all the row. Thank you so much.

S1 **Speaker 1** ▷ 16:09

Any other questions guys? Okay, perfect guys. So let's actually work on our first, we have an activity, our first activity for c plus. And I wanted to create it like you want, just want this one and that you see 7,895 rooms and two colors and you get it like this.

S3 **Speaker 3** ▷ 16:50

Wait, how do you do that?

S1 **Speaker 1** ▷ 16:52

Just remove the head here.

S3 **Speaker 3** ▷ 16:55

Oh, okay.

S1 **Speaker 1** ▷ 16:57

Okay. If I put head here again, I put it like that. I can bring 10 values for example.

S3 **Speaker 3** ▷ 17:03

Oh, I see.

S1 **Speaker 1** ▷ 17:04

I can bring a hundred values. UT hundred is not printing the hundred. I think 50, 50 years is gonna do 50 years.

S3 **Speaker 3** ▷ 17:17

Why didn't print a hundred? Oh because there is no a hundred in the database.

S1 **Speaker 1** ▷ 17:22

No, it's hundred. But in Jupyter Novo it's not going to show me the hundred.

S3 **Speaker 3** ▷ 17:27

Oh, I

S1 **Speaker 1** ▷ 17:28

See, okay. It mean like the street. You see it. That means that going okay. Oh hey Hugo, the street name's still there because it's considered the index, right? Exactly, exactly. Yep.

S1 **Speaker 1** ▷ 17:49

Yeah. Okay, perfect guys. Oh, if we go here, do a data analysis with pandas and number two activities, we're gonna do lock with good movies, okay? We have instructions here, we have a resource, we have a list of the columns in the data set. We are only interested in IMB data, so create a new table, blah, blah, blah. And finally export this file. Gonna stretch it. That's, that's a part that I was talking about that where you actually can see it more clearly. So we have 15 minutes to work in this activity. Guys, please start working. If you have any questions, we can open the breakout rooms and we are actually 4 48. So I'm gonna open four breakout rooms and you can go and ask questions in the breakout rooms before the activity finishes. I'm gonna close the record rooms so you can come back and see if you can stay here and ask questions here. That's perfectly fine. So we have 15 minutes to work in these guys. We start working and then we can review Hugo, sorry, which activity again? Number two, activity number two s St. Good luck. Okay, thank you.

S4 **Speaker 4** ▷ 27:26

Hey Hugo, quick question.

S1 **Speaker 1** ▷ 27:33

Yes? Yeah. Yes.

S4 **Speaker 4** ▷ 27:35

In line seven or in the input seven, when you give DF lock and you give the call in, what does that specify exactly?

S1 **Speaker 1** ▷ 27:47

Line seven, let me check right now because I closed it.

S4 **Speaker 4** ▷ 27:51

Okay, so you're using lock to select all rules from street food name up until and postal community

S1 **Speaker 1** ▷ 28:03

Nine, seven. Yes, exactly. I am selecting

S4 **Speaker 4** ▷ 28:06

Is that just the syntax when you put the column in before the comma,

S1 **Speaker 1** ▷ 28:11

The column before the coma, that means that select all all the roles. Ok, okay. And then comma, and then you specify which columns.

S4 **Speaker 4** ▷ 28:21

Okay, so if you don't, and you've specified head, so it'll only give out the first five.

S1 **Speaker 1** ▷ 28:27

Exactly.

S4 **Speaker 4** ▷ 28:28

And the same thing with IOC is the indexing with the numbers and not with the names.

S1 **Speaker 1** ▷ 28:36

Exactly, exactly.

S4 **Speaker 4** ▷ 28:38

Okay, got it. Thank you.

S1 **Speaker 1** ▷ 28:40

No, five more minutes guys. So maybe we can start checking.

S1 **Speaker 1** ▷ 35:14

Perfect guys time up. So we're gonna, we're gonna actually start, okay, we're gonna, I'm gonna start sharing my screen, please let me know if you can see my screen with a thumb up. Perfect. Okay, so good movies, okay, in good movies. First we're gonna read from resources, the movies score db. Okay, then we, we read with pdsb and then we, we, we, we print the head just to take a look of this huge table. Okay? Now I am going to print list all the columns in the table. So basically you have to put movie file column and here's gonna bring all the columns that you have in that tape. Okay? Then in the, in the, in this instruction they tell you we only want this columns, IM db Okay, so I am DB is here. Where else Im DB here I am DB here, right? We have IMDB here as well. V count and we're all, that's it. So it's four columns, eh? So I'm gonna bring that Im DB data frame movie five data frame, double double brackets here. And then I am going to put film, I am db, I am DB norm IMDB norm round and I am DB user. Okay, five, five columns. I print my head here I can see an overview of my data frame, my new data frame.

S1 **Speaker 1** ▷ 37:22

Now they ask us to do some filtering. We only like good movies. So find those that score over seven and ignore the norm rating. Okay? So ignore, ignore the non rating, these two and then find the IMD score higher than that. So basically you have to put the good movies df, create a new variable that you're going to save your data frame call your movie file data frame lock, okay? Because we are just going to take from the columns and this is a good example. You're gonna use lock when you want to refer to the columns because the columns, they are almost every time strings. Okay? So now you can do movie file D F imd, imdb, sorry, bigger than seven. This is, remember it's going to be my rope. This is filtering all my ropes, okay? And which columns I am going to take?

S1 **Speaker 1** ▷ 38:30

I am going to take film, I am db am DB user. Both. Why? Because it says ignore norm rating. Okay, so after that he is asking us to put fine less popular movies tools with fewer than any cables. Okay? I am going to do exactly the same, but now I am going to filter by imd, user bot both count. And I am going to bring again the same film, imdb imdb user bot count. And then this is my finite table. Finally export this file to a spreadsheet so we can keep a track of of how new features watch list without the index. Okay, so index is gonna be fold here cause it's telling us, and I'm going to put in my unknown movies Excel. I am going to create a folder here that is gonna call output. And inside this folder I am going to create a new file that is with an extension of Excel sheet. What will be the is if I want to export to a csv, what will be the writing, what will be the function, the command

S3 **Speaker 3** ▷ 39:52

To underscore csv.

S1 **Speaker 1** ▷ 39:54

Exactly, yes. And guys, any questions

S3 **Speaker 3** ▷ 40:02

When I have a question and the last, in the last line this it says export the file to us spread so we can keep track of the new FU future watch list. Does that specifically has to be Excel or it can be in a different,

S1 **Speaker 1** ▷ 40:23

We got CSV as well.

S3 **Speaker 3** ▷ 40:25

Oh that okay. Okay, got it. Thank you.

S1 **Speaker 1** ▷ 40:28

No. Any other questions? Okay, perfect. So we're gonna jump now to cleaning data. Cleaning data guys. So basically cleaning data when dealing with Maci massive data sets, it's almost individual that we will encounter duplicate roles in consistent spelling and missing value. This is hundred percent real price. So we're gonna start doing some delete. For example, we're gonna introduce to delete columns, counts to count specific data in in Rose we're gonna use a drop Now drop Now to eliminate new values, we are gonna use value counts to count whatever we have in a specific column. And we are gonna use replace to replace whatever names we have in that specific column as well. So let me show you now my, okay, cleaning data. So I read my and read it here, printed here, df, right? I can see name, employer, city estate, sip amount and memo cd memo CD contains like any values.

S1 **Speaker 1** ▷ 42:11

So this is a weird column, so I'm just gonna get rid of it. How I'm gonna get rid of it, I'm gonna include, use Dell DF memo, just like that. That easy. Okay, I print my new data frame, I don't have the column anymore and then I am going to count all my values in all my columns. So I can see if there is any inconsistency in my roles. I put BF just like that and I see name 2000 values employer 18,000 values. 19,000 values 19. This is inconsistent, right? Must have missing information here because why I name if, if it's a, if it's a table with the same number of columns and same number of rows why I in name, I have 2000 values and employer I have 18,000 values. I am, I have missing values. So I want to delete all the blank spaces that I have.

S1 **Speaker 1** ▷ 43:09

Okay? In order to do that, I, I just gonna put F drop na how any, if it's black, if it has na, if if it has a space character there, whatever, just delete it, okay? Drop it. So if I count, I see that my table now contains the same number of values. Okay? Of course guys, this is something that whenever, for example in row one I have an empty cell and then that belongs to my column one. That empty cell is gonna disappear, it's gonna drop it, okay? And it's gonna count it. I mean it is not going to count it, but it's not going to disappear it. Okay? Why is this? Because I want to keep the same, the same size of my table. Okay? That's clear because I don't want to have an empty space and then a call another cell. That doesn't make sense, okay? In my rows and columns, that's why dropping is not removing that cell. No, I'm gonna leave that cell alive, but I'm gonna remove whatever value is easier. Okay? So the machine or my, my python doesn't count that value. I don't care if it's an na, if it's a zero, if it's don't count it, okay?

S4 **Speaker 4** ▷ 44:54

You're not going to delete the information, you're just going to omit it from the count.

S1 **Speaker 1** ▷ 44:58

That's it. Exactly, exactly. Got it. Thank you. I'm not dapping rose because sometimes they ask me, but why if you're dapping rose, your table is gonna be over. Like, but no, I'm not disappearing. Okay,

S6 **Speaker 6** ▷ 45:12

Quick question. When you, sorry. When you do that, then if you take the mean something like that, it's also not any functions you are using, it's not including those drugs as well.

S1 **Speaker 1** ▷ 45:23

Exactly Jeff. And that yes,

S7 **Speaker 7** ▷ 45:29

In that case if, if you write it as is your excel, will it include those missing values? Like deleted values or excluded?

S1 **Speaker 1** ▷ 45:40

You are not going to include those values as values, but you include the cells, okay? The empty cells, like the little square cells,

S7 **Speaker 7** ▷ 45:50

The empty cells is gonna be there if you, if he,

S1 **Speaker 1** ▷ 45:53

They're gonna

S7 **Speaker 7** ▷ 45:53

Be there after, after cleaning.

S1 **Speaker 1** ▷ 45:55

Yes, exactly. Okay.

S1 **Speaker 1** ▷ 45:57

Okay, perfect. No problem. Now if I put F, this is important because you, you are going to find a lot of errors when you are working and then you forget, oh, I forget the date of what I'm reading. For example here sip, it's normal that it comes like an integer or float, right? But to be honest, you can't work with a sip that is an integer. You will have to work with a C value that it's an object because you want to find a specific C for example, 79, 9, 12, right? And and if you put it as as an integer, you're gonna try to find numbers instead of letters. So that's why zip you normally handle it as a string like an object. And here you see it like they send it a number, okay? So in order to change that zip you will have to put F as type function parentes app, parentes C, my sip column.

S1 **Speaker 1** ▷ 47:07

Yes it's a string. The same thing if you want to change it to float, if you want to change it to a bullion, this error race, error race basically is handling my errors. So if you want to handle the error when a user actually put your string as an integer, it's gonna be handled by this part. Why this is important and we're gonna talk about error handling, don't worry about it, but it's important because when you want to get an input for a user for example, and the user types an integer and is not type and string, then this is going to handle that or without any conditions, okay? You don't have to put, if the user puts an integer then send this message. No, it's gonna handle it by this part, okay? This can be rice or not rice. One thing guys, I want you to, this is not clear for me, I still don't know what's going on.

S1 **Speaker 1** ▷ 48:11

Okay, go here then just put at type pandas, go enter and you have this great grade documentation pandas that gives you everything about the function that you're using. A type is expecting the type copy name errors, but what is errors? Expect errors. You can have two pins, you have the rise, ignore, or the default that is right, okay? And you have the rise, ignore, and the default is right. Default means that if you don't put errors inside, for example, this one I didn't put copy, my default is true and it's controlled raising of exception on invalid data or provided data type and they give you actually how to do it.

S1 **Speaker 1** 49:08

Okay? So it is really, really useful. You're gonna be learning a lot from here as well. Delete me and how do you find it? Just like I type anything. Okay? So whenever I verify my C, the ffc, the type is gonna be a no because it's an object. Okay? So I'm gonna display an overview of the employer's column, the employer, the value count. I'm gonna see that I have several things in internal values in that specific column. I have not employed none self-employed. I have another one that is called self, like this self can it be maybe the same as this one? So I can see something weird here as well, right? As well. The names I don't like it a lot. Not employed non was meaning non, right? So the first thing is that I'm gonna put together these two because self-employed and self for me seems the same. Okay? So I'm gonna do df, employer equal two DF employer and then I'm going to use the replace foris for entities. This is a format of of a dictionary, okay? I'm using brackets. They currently brackets sales sell replaced by self-employed and self-employed. It replaced by self-employed. So as soon the machine knows that this tool, the same name is gonna sum up the value here. Okay, so verify the cleanup DF employees value account. So now I just self employees, sorry, yes, I

S4 **Speaker 4** 51:04

Have a question. So what the moment you change self to self-employed, isn't that enough? Why are you again specifying self-employed? As self-employed?

S1 **Speaker 1** 51:14

Cause I have two categories as you can see here. This one does not have any.

S4 **Speaker 4** 51:18

Oh, okay. Cash in the, okay, got it. Alright, thank you. Sorry.

S1 **Speaker 1** 51:24

And that's it. I'm gonna rename the not employed as well if employer that replace not employed by unemployed right now it looks like this, of course you can keep going, you can keep doing things right. Then I want to display a whole statistical overview. This one, you know it, you know how to do it. The F described and it's gonna click only the in to give you the count. Mean standard as deviation, minimum maximum value on the percent. And I want to put this into a csv, the F to CSV encoding, you can use it or not. And as well, again, if I don't understand, I go to my documentation question. So cleaning guide,

S3 **Speaker 3** ▷ 52:10

I do, how do you know what self and self employer is the same? Where do you see that

S1 **Speaker 1** ▷ 52:17

Basically? Yes. That's a great question, Denise. Basically I am assuming that because of the data that I am reading here, okay, I am doing that. So, but you can, you can check the original, see it by yourself.

S3 **Speaker 3** ▷ 52:38

Okay.

S1 **Speaker 1** ▷ 52:38

Okay,

S1 **Speaker 1** ▷ 52:41

No problem. Any other questions? Okay, perfect. So for this activity, activity, Hong Kong LPG appliances, we have 20 minutes actually because we are going to work in cleaning our data. In this activity guides, you're gonna see that it says p r cleaning appliances data. The main purpose of this activity is to get groups of your classmates, get together with the groups and then start sharing your screen and someone else help you. Okay? Again, it's not mandatory. I am going to open the rooms. If you want to go to a room and start sharing with your classmate, that's totally fine. If you feel more comfortable working by yourself, that's perfectly fine. Okay, so 20 minutes guys, I'm gonna open the breakout rooms again and I'm gonna give you like a notice when five minutes start missing. So you can go to the room, you can stay here. But let's start working on this activity. Yes, you have questions. Make sure to ask

S3 **Speaker 3** ▷ 55:26

I Google question. I was assigned on room four, but I don't, I don't think no one, I mean it's Dan, but I'm calling him and I, he's not there. I'm not sure if I can switch.

S1 **Speaker 1** ▷ 55:37

Yes, totally. Denise, let me switch you to another room. No problem. Denise, let me check who is on. I can switch you to room number seven.

S3 **Speaker 3** ▷ 55:49

Okay, cool. Thank you.

S1 **Speaker 1** ▷ 55:51

No problem.

S3 **Speaker 3** ▷ 58:58

Sorry. Well this is a very silly question. I, I'm, are we working on exercise five?

S1 **Speaker 1** ▷ 59:08

No, it's four.

S3 **Speaker 3** ▷ 59:11

Four. Okay. Ugo, can I ask you real quick, I know that you mentioned how to get the path, but I don't see it in my Mac. Can I turn my screen real quick?

S1 **Speaker 1** ▷ 1:00:46

Yes, please. And

S3 **Speaker 3** ▷ 1:00:48

So this is, I know yours have like some something here. So I click on here and that's not it. I click here and either

S1 **Speaker 1** ▷ 1:01:05

Number two go to your left, I mean to your right No, but Right, right. Yes, you have it there in,

S3 **Speaker 3** 1:01:16

Okay, but how can I copy the path?

S1 **Speaker 1** 1:01:19

Oh, the path in Mac.

S3 **Speaker 3** 1:01:21

In Mac. Yeah.

S1 **Speaker 1** 1:01:23

Lemme check because in Mac

S8 **Speaker 8** 1:01:27

It's, it's if you right click on it or control, click on the path or on the file.

S3 **Speaker 3** 1:01:35

This one?

S8 **Speaker 8** 1:01:36

Yeah, you do control click it should be control or like and then get Info.

S3 **Speaker 3** 1:01:49

Info.

S8 **Speaker 8** 1:01:55

So you see where it says where,

S3 **Speaker 3** 1:01:58

Where. Yes.

S8 **Speaker 8** 1:02:00

So if you copy, try copying everything after. Wear up until activities and pasting it somewhere and seeing how it pays. I think sometimes it pays correctly.

S3 **Speaker 3** 1:02:09

So this and start at zero four activities, right? Or

S8 **Speaker 8** ▷ 1:02:13

Oh, start at Macintosh? Yeah, just start at the top.

S3 **Speaker 3** ▷ 1:02:17

Okay.

S8 **Speaker 8** ▷ 1:02:18

And then just try pasting that somewhere and seeing how it pays.

S3 **Speaker 3** ▷ 1:02:24

Let's see. It works. How can I check if it works? I just run it.

S8 **Speaker 8** ▷ 1:02:31

So you have to put strings around it.

S3 **Speaker 3** ▷ 1:02:35

Put I'm sorry, put what?

S8 **Speaker 8** ▷ 1:02:37

You have to put quotations around your path.

S3 **Speaker 3** ▷ 1:02:43

Oh yeah.

S8 **Speaker 8** ▷ 1:02:48

And if you're trying to read it, you'll have to try and read it in. But right now that's just your path to that directory. It's not gonna be the path to the actual file name.

S3 **Speaker 3** ▷ 1:03:10

So I'm confused. So then what else should I do?

S8 **Speaker 8** ▷ 1:03:15

If you want it to your actual file name, you could do the exact same thing you just did, but right click on your file name and get info or if you know where it is from in there. So if you go to activities, to the very right hand side of activities

S3 **Speaker 3** 1:03:32

Here.

S8 **Speaker 8** 1:03:33

Yep. And then you do forward slash and then tab.

S3 **Speaker 3** 1:03:44

It does not, it doesn't work.

S8 **Speaker 8** 1:03:47

Oh, okay. It's not auto filling for you?

S3 **Speaker 3** 1:03:50

No. So I guess, let's see, it's on Unsolved.

S8 **Speaker 8** 1:04:00

So

S3 **Speaker 3** 1:04:01

Is that unsolved?

S8 **Speaker 8** 1:04:03

Yeah, I don't, whatever you're trying to read in, you wanna, you wanna read in the actual CSV name?

S3 **Speaker 3** 1:04:12

Okay, so it's unsolved but this is not a csv.

S8 **Speaker 8** 1:04:20

Yeah, that's the, that's just the notebook file that you're in.

S3 **Speaker 3** 1:04:26

I'm very confused.

S8 **Speaker 8** 1:04:29

Try resources

S3 **Speaker 3** 1:04:34

Also. That'll be in resources, this one. So it'll be then fold folder resources. So did I need all this though? Or I just put resources?

S8 **Speaker 8** ▷ 1:04:49

Depends, it depends where you are right now. If you're in that unsolved, this is, if you have no idea where you are, if you, if you know where you are, then you can just navigate from wherever this notebook is located, which I think is just one directory up. So you don't need that whole thing. You can just do period, period slash resources slash and then the name of the CSV file and that that will be the path to where you're, but if you have no idea where you are, then you can do it that way.

S3 **Speaker 3** ▷ 1:05:20

Got it. So resources and then dot, is that

S8 **Speaker 8** ▷ 1:05:24

Not at dot? It's a slash.

S3 **Speaker 3** ▷ 1:05:27

Oh that's right. Four slash and then unsolved and then resources.

S8 **Speaker 8** ▷ 1:05:36

No, unsolved just the name of the CSV file. Got it. Cause the CSV file is and resources.

S3 **Speaker 3** ▷ 1:05:41

Oh I see, I see what you mean. Okay. This one. Got it. Yeah. Awesome. Thank you. Well lemme see if it works though.

S8 **Speaker 8** ▷ 1:05:50

Yeah, right now you have a couple of extra quotes too that you don't need. Basically you always want that to be in red and you see how it turned black there. It means that you've closed off your quotes

S3 **Speaker 3** ▷ 1:06:05

Like this.

S8 **Speaker 8** ▷ 1:06:06

Get rid of the, yeah, that one. There you go. Oh, but keep the first ones.

S3 **Speaker 3** ▷ 1:06:13

Oh the first one Cuz always gonna take, yeah, I got in

S8 **Speaker 8** ▷ 1:06:15

And then delete that second one. Yep. And then just close it off when you're done typing the CSV path.

S3 **Speaker 3** ▷ 1:06:22

Okay. Pta, lpg. And then close and like that.

S8 **Speaker 8** ▷ 1:06:34

Yep.

S3 **Speaker 3** ▷ 1:06:35

Awesome. And got it. All right, thank you.

S8 **Speaker 8** ▷ 1:06:42

Yep.

S9 **Speaker 9** ▷ 1:06:43

Denise?

S3 **Speaker 3** ▷ 1:06:45

Yes.

S9 **Speaker 9** ▷ 1:06:45

Can I share one trick that I do?

S3 **Speaker 3** ▷ 1:06:48

Sure, please do.

S9 **Speaker 9** ▷ 1:06:50

Can you go to the file and then do save as

S3 **Speaker 3** ▷ 1:06:55

The file and then save before, was that the path is already there?

S9 **Speaker 9** ▷ 1:07:01

No, regardless of that you can figure out where your I P Y N file is saved and from there you can select what path you need to direct to.

S3 **Speaker 3** ▷ 1:07:10

Oh I see, I see. Denise can you, can you please share your screen cuz

S9 **Speaker 9** ▷ 1:07:14

I have a Mac too,

S3 **Speaker 3** ▷ 1:07:15

I don't follow. Yeah, yeah. So let me, lemme go back. You're, I'll go to file, see that

S9 **Speaker 9** ▷ 1:07:25

You can see you are an unsolved directory from and the path you can go, you have to go up. So you put double dots and then resources and then the file name CSV

S3 **Speaker 3** ▷ 1:07:37

Here.

S9 **Speaker 9** ▷ 1:07:39

No, you don't need to write, you just need to figure out where you are. So from here you can find out where your file is saved.

S3 **Speaker 3** ▷ 1:07:44

Oh yeah. And

S9 **Speaker 9** ▷ 1:07:45

All the class activities, I think it is the same way.

S3 **Speaker 3** ▷ 1:07:49

Got it. So once that I know where I'm at, I can just see where the CSV file is that

S9 **Speaker 9** ▷ 1:07:55

Yeah, you can just go one folder up for double dots and then the same the thing, the same thing that you did right now. But for each and every class activity like this, you can find out where your IP by n b file is stored.

S3 **Speaker 3** ▷ 1:08:09

Awesome. All thank you so much. You're

S9 **Speaker 9** ▷ 1:08:13

Welcome. Thank you.

S8 **Speaker 8** ▷ 1:08:14

Yeah, and I'll share something in the chat that this is a way to, in Python, be able to figure out what directory you're in. So that will save C W D is just getting your direct directory that you're in at the any given moment.

S2 **Speaker 2** ▷ 1:08:29

There's one more thing Dennis that I did. I just copied the CSV file and pasted it into the unsolved version and it worked for me. So if you want

S3 **Speaker 3** ▷ 1:08:39

You did, you did what? Sorry what? You went to the unsolved folder?

S2 **Speaker 2** ▷ 1:08:44

No, so you go to the resources in your tech desktop, copy that C SV file, paste it into the unsolved version,

S3 **Speaker 3** ▷ 1:08:54

Like basically see resources.

S2 **Speaker 2** ▷ 1:08:57

Yeah, copy that. Open that folder.

S3 **Speaker 3** ▷ 1:09:01

It's open.

S2 **Speaker 2** ▷ 1:09:02

Okay, wait, I don't see that. Yeah, yeah. So copy that CSV file and then go to Unsolved.

S3 **Speaker 3** ▷ 1:09:15

Unsolved.

S2 **Speaker 2** ▷ 1:09:17

Yeah. Paste it here. Now go to your code Jupiter. Go to your Jupiter. Now, now in your CSV file is equal to part in bracket. Just type the, the, the file name DG LPG csv. That's it.

S3 **Speaker 3** ▷ 1:09:50

Sorry, where I go? Where I go here on the doc?

S2 **Speaker 2** ▷ 1:09:53

No, no, no. You have on the left hand side your code. Right. So in line, in line number two, delete all through resources. Delete all the all the, yeah, yeah, keep that the selected version and delete tell resources forward slash

S3 **Speaker 3** ▷ 1:10:14

Got it.

S2 **Speaker 2** ▷ 1:10:16

Yeah. Clear all. Yeah. Even double dots

S3 **Speaker 3** ▷ 1:10:22

And then paste.

S2 **Speaker 2** ▷ 1:10:24

No, you are done. And now in the second line you just read that file. So how you do it right? Oh

S3 **Speaker 3** ▷ 1:10:31

Wow. Okay. I see, I see.

S2 **Speaker 2** ▷ 1:10:34

Oh that's,

S3 **Speaker 3** ▷ 1:10:35

That's great. Thank you guys. My team. Woohoo. You thank you so much. Yes, yes I did. Thank you guys.

S2 **Speaker 2** ▷ 1:10:49

Thanks.

S1 **Speaker 1** ▷ 1:12:01

In two more minutes guys, I'm gonna start showing the answer. Okay. Perfect guys, welcome back. I'm start sharing the answer here and please let me know if you can see my screen thumb up if you can. Perfect. So here is it the HK L pg, that's the one that you work the cleaning. Apply appliances. This one. So we read that E S B E A L P G. So the first part that they ask us to do is, is to count columns, right? Look for missing values. So LPG reviews, columns, dot comt, concrete, no, no resource. Just missed one. Yes. This one reduce the columns that are in English. Okay, so we are gonna reduce for this column, part type, brand model, other information, place of manufacturer, click and telephone number, approval expire date. Okay, I have those columns already. Then I'm gonna count all the values in my column and I can see well that everything is not because of this one and 9 36, I'm gonna drop now the reduced columns.

S1 **Speaker 1** ▷ 1:15:35

Okay? So when I print it, everything is printing correctly, okay? After that I'm going to use the list of unique values of applicant to locate any data may be the same. So no new, that's the name of my new data. Frame A then unique, right? And it is gonna show me the unique names. Now that I have the unique names, I'm gonna combine similar applicants together. So in here I can see applicant, right? And I can check like which ones are similar. So I'm gonna use a replace, replace crown gas stop holding company limited to crown gas stop Ltd and P limited. Okay, I'm gonna replace the names on those ones in order to get my values or see my values more understandable then I am going to do exactly the same thing. I'm gonna check to see if you combine similar applicants correctly in applicant. So applicant dot unique again and then you look into this data in order to see if what you did here, it's working fine. Okay, one example here.

S2 **Speaker 2** ▷ 1:17:07

Sorry, you got to interrupt. So is there only way to look at the data and find out the similar values or is there any other way to find out automatically that the code should give me which values are similar? I mean this is like a pretty much, yeah, even this is a big data. My So how, how would you find out?

S1 **Speaker 1** ▷ 1:17:35

Yes, right, you're right. Basically this is like just the, the, the part on how to use this unique, how to use the replace, those kind of things. But we are gonna see specifically that project where we actually can see and compare data that is in the same column. Okay. And we can see these two are similar, these two are equal. These, these are different and they give you how many are equal, how many are different. But that's later on. Okay, but we're gonna see

S2 **Speaker 2** ▷ 1:18:11

Got, thank you.

S1 **Speaker 1** ▷ 1:18:12

No problem. What else? Then we are gonna create a new data frame that looks into a specific base of manufacturer. Basically I'm gonna use a lock with my no new L P D D F place of manufacture. It's equal to mileage. Okay? So it gonna filter to mileage. Then I'm gonna create a new data frame that looks into a specific place of manufacture again, but it is gonna be Portugal least. Okay, so I have Malaysia and I have, and that that was it. Any questions? Yes.

S9 **Speaker 9** ▷ 1:18:50

How do you get that scroll Barb? With all the results? Usually it shows that three dots, right? Is there anything different here?

S1 **Speaker 1** ▷ 1:19:01

Well, it's nothing different I think is my version. I didn't do anything different in my coat or anything like that. But normally when you do these kind of filters you get that scroll bar. If you are not getting it, we can check you. You, you have the same problem with printing something else, right? Last class maybe is the version on Jupiter?

S9 **Speaker 9** ▷ 1:19:27

No, I think I have the most latest one.

S1 **Speaker 1** ▷ 1:19:30

Okay, so we can, we can check on that.

S2 **Speaker 2** ▷ 1:19:33

Okay. Okay. Thank

S1 **Speaker 1** ▷ 1:19:35

Any other questions here guys?

S1 **Speaker 10** ▷ 1:19:37

Yes. Yes. Hugo, can you just scroll up to the in number two or number three I think

S1 **Speaker 1** ▷ 1:19:45

Yeah,

S1 **Speaker 10** ▷ 1:19:48

Fine. Yeah for the cleaning. So yeah, for the, this intersections cost, like what I see on the, on the PowerPoints at say we need use the d e A orders so we can, we can delete the all non-English one. So is that okay to using in here?

S1 **Speaker 1** ▷ 1:20:13

Totally change? Yes, you can use that as well

S1 **Speaker 10** ▷ 1:20:18

Or we need to still using the likes with the written?

S1 **Speaker 1** ▷ 1:20:22

No, you can use the delete functional.

S1 **Speaker 10** ▷ 1:20:25

Okay, let's, let's also work for the, for this kind of functions, right?

S1 **Speaker 1** ▷ 1:20:31

Yeah.

S1 **Speaker 10** ▷ 1:20:31

Okay. Okay, thank you.

S1 **Speaker 1** ▷ 1:20:33

No problem.

S1 **Speaker 11** ▷ 1:20:35

Hugo, can I see your line too please?

S1 **Speaker 1** ▷ 1:20:46

Thank you.

S1 **Speaker 10** ▷ 1:20:47

And and also Hugo, can I just ask on the line two the last line with I pg, you know scratch df, what's that mean? Cause you already just defined LPG slash slash df. So what's the last LPG slash DF mean?

S1 **Speaker 1** ▷ 1:21:07

This is just to bring the table.

S1 **Speaker 10** ▷ 1:21:09

Okay, don't, okay, got it.

S1 **Speaker 1** ▷ 1:21:12

No problem.

S3 **Speaker 3** ▷ 1:21:14

I got a very silly question for the type one. Type one. Is that how the database come? Like there's no way to have those characters, translators or anything. So if I run into that, like how am I gonna know what that means?

S1 **Speaker 1** ▷ 1:21:29

Yes. No, maybe you just have to delete them. That's why we are deleting them.

S3 **Speaker 3** ▷ 1:21:34

Okay. God,

S1 **Speaker 1** ▷ 1:21:36

No problem. Okay, perfect guys. So we're gonna go to the next part of the class that actually basically just going to, to a review. We're gonna do a review of what we have seen so far as recap. You can do it with me, you have it here in GitLab as well. And we're gonna talk about this part, okay? So again, I'm gonna start again. Import pandas library, import pandas as pd. Then I'm going, I'm, I'm going to create a CSB path resources, the name of the file that I want to read. And then after this we are going to, I think yes, we are recording, right?

S3 **Speaker 3** ▷ 1:22:24

Cool.

S1 **Speaker 1** ▷ 1:22:25

Yes, no. So then I am going to read that SVL pd read sv and then pointing to my CSV path. Then I'm going print the first five rows of data all to the screen fires the, and then it give you a warning. You're gonna see this warnings quite a lot. And this warning is actually because it's identifying data type warnings like for example, mixing types on c mixing types on that Jupyter, understand that a sip must be an an object and not an integer kind of things. And it gives you warnings. Okay? Don't worry too much about warning, you can them, I mean it, it is possible. But for example, these kind of things, low memory, it, it's, it's setting that you have to do on in Jupiter. But I will say right now don't worry about this, okay, just keep, keep working on Jupyter.

S1 **Speaker 1** ▷ 1:23:29

So after that I'm gonna do account, the account seems odd. You can see here I here, I actually before doing the the the drop, now actually here we are renaming the columns, okay? Renaming a specific column, how columns equal to and then with type eight given or received code to this name and property loss to this name, okay? Because it's property and it's property then it's gonna give me reduced to condoms, reporting gear, fire department name, incident date, incident type. And they give me all these columns, right? So I just copy and paste it here so I don't have to type all of them. You can do this kind of thing in no plus plot or in an editor. Visual video code maybe can help you to do that. You don't type like each column one by one. Okay? Then I print the head here and then I'm going to actually fill an A.

S1 **Speaker 1** ▷ 1:24:44

Remember I drop an A, I'm going to fill an A. This is going to help me to put for example, the action taken one, fill it with an empty space, the action taken two, fill it with this space here guys, one important thing, I'm just following instructions, okay? This depends on how, what you're doing with your table, what you're doing with your data prep, what you want to do and that those are the instructions that you're gonna do. Here we are just giving you like an A drop, an a rename, replace like those functions. So you can think of one of those faster and then you see how they work. Okay? For example, replace other fires with zero and then specific things that these structures are, these functions are instructions are given. Okay, drop. Now again, I am going after I replace everything, I'm gonna drop now help any and then counting and everything.

S1 **Speaker 1** ▷ 1:25:49

Look metric now and then I'm going to start doing both. Filter three, okay, log fires, cleaning the F, property loss bigger than zero and fire cleaners the F content loss as well. Bigger than zero, I bring all the colors, okay? I am going to do the value count in order to see if something like this is repeating or if I can actually replace something. And I am going to convert the the C count series into a data frame because these city count is a series as you know it's just a column. So I am going to convert PD data frame C counts, OK C counts. Once it's converted I can see it like with this index. So now that I have the indexes, I can actually for example, convert the column name into zoom of loss incidents. So city loss column, the F is gonna be equal to my data frame, my same data frame, but renaming it column again dictionary typing incident to zoom of loss incident. Then I am going see the first five rows and I am going do the now. Okay, I'm gonna calculate the number of death from fire incidents where loss occur. So that equal to loss. F taking the log, the F data frame that is here. Okay?

S1 **Speaker 1** ▷ 1:27:28

And I am going to take only fire service debt, some plus log, the other fire dead data types. I'm gonna see the data types here. Object flows. So I want to change the alarm date and time, the arrival date and time and the last unit, clear date and time for daytime. Cause they, they can be, they can be on the objects, they have to be date times and then I can see it here. Okay? Now I can find the response time in seconds. So I can do some timing mats here. Loss response, time in seconds. Loss, the F is gonna be arrival. Date and time minus loss. D, f alarm, date and time. When I soon extract divide, multiply time data types, I can do it here like with simple maps. Okay, once I have that, I am gonna convert my response time in seconds into dot dt. Total seconds. Okay, and again, this is gonna be a, a Skype. I am going to convert this type integer, okay, because I don't want to see the seconds anymore. I want to use the seconds as an integer.

S1 **Speaker 1** ▷ 1:28:58

Once that I have that I can check the data for columns of your choice. I can plus the F and then just put columns that I'm interested in and I can check it out here. Okay? This file guys is gonna be available for you, like right now if it's not dry, now available for you. So you can check it out, you can play with the, with variables, you can delete variables, you can see how the data frame is printing, you can remove columns, you can put new columns around here, okay, so you can check it out. Questions so far guys? On each, on any, on any specific function here.

S1 **Speaker 12** ▷ 1:29:45

Hey Hugo, can you scroll down to where you have changed the type the data type or

S1 **Speaker 1** ▷ 1:29:55

Here

S1 **Speaker 12** ▷ 1:29:56

I'm just, yeah, I'm just trying to figure out date time. 64. Does the 64 have any significance or is that just a, a general data type for daytime?

S1 **Speaker 1** ▷ 1:30:06

It, it's a general data type for daytime. Okay, thanks. No problem. Any other questions? Okay, perfect guys. So let's take a 15 minute break and then we can come back for the second part of the class guys. Thank you. Going into the toilet, perfect guys breaks up. We are actually going to start checking group by. Okay, I'm gonna start sharing my screen and let me know when you can see my screen. Thumbs up? Yep. Okay, perfect. So let's group by group I, it's like the group I in Excel to be honest. Method for filtering data.

S1 **Speaker 1** ▷ 1:45:56

So you're gonna see here that to split the data frame into multiple groups and group by state, we use DF group by, and then the columns, the group by method returns a group by object that can only be accessed by using a data function on it. Why is a data function zoom count, all that kind of things. Okay? For example, in this example group, ctd F equal loss, DF, group by incident CT incident C, print the group by C tdf, but group by CDF dot count, okay? And it's gonna count, even though that doesn't make sense, it's gonna count. Okay? So for example, the data frame metric that makes possible to create new data frames by using only group by data here data frame can also be created by selecting a single series from a group by object and passing it into in as the value for a specific call. Okay, we're gonna see here the example save loss soon as series 80, property loss grouped CDF property loss and grouped CDF content. Lo content loss. Okay, we're gonna see this in the code, don't worry about it.

S1 **Speaker 1** ▷ 1:47:30

Yes, you can multiply, you can do a multiple columns to be honest, just two because once you do it three by three columns, like incident, city incident type code and another one, it is not going to do anything. It's just gonna stay in the second column that you're grouping by. Okay? It's gonna show you the third column, yes, but you're gonna realize that it's not grouping anything, so it stays in only two. Okay? You can create a new data frame from a group by and yes, let me show you the example here. Basically we have, I'm gonna rename these columns, property lost, this is the one that we just went through the time, total, total seconds, but we're gonna go to the group file, okay?

S1 **Speaker 1** ▷ 1:48:30

Value counts, okay, here's, I'm gonna start grouping by group. Ttd, F loss, df, group I, incident C, okay, incident C, basically it's gonna put my incident here. The column is gonna group by incident CT and I am going to count, okay? The, the same example that we use, you're gonna see that it counts 1, 2, 6, the count is gonna be the same, okay? Because it's only counting values in each of the, in each of the columns, okay? It's not counting each here sum anything. So it is just count. So I'm gonna group by I I, I'm just gonna, sorry, I'm just gonna take only two columns with my data frame that is already grouped by, remember by incident and I'm gonna do a sum here and now it makes sense. Okay, property lease in am Amazon 65,000, sorry, content loss in am Amazon as well, 5,000 and so on.

S1 **Speaker 1** ▷ 1:49:44

Now I am going to do, I am going to save lawsuit sums as a series. Okay? So this exactly these two, I am going to save save it separately as a series. As as, as a separate columns, okay? Group by C and I just put it here, group by c.com, sorry. And count counts as well with the thumb as well. Okay? So I have two series here, one that is the content and the other one that, that is the property. Okay? And now I'm gonna create a new data frame using the count and those amounts so I can use them because it's a series, it's not a column, it's a series. So I can use that date. Okay? So city summary is gonna be one to P data frame, number of lost incidents and I'm gonna use the city lockdown, total property laws, city property loss and content loss, city contents loss. And I can see that here with all three together.

S1 **Speaker 1** ▷ 1:51:00

I can do another group by, by two different columns group by incident city first. So incident city is gonna be to the very left. And then after that incident type code that is gonna be to the right, okay? And you gonna see for example is actually grouping by incident city, but then incident code 11, 11, 11, 1, 11 3 11, 4 31 you see and will arrive from a new category, okay? I can convert a group by object into a data frame. Yes, this group by I can convert it into a date frame group loss incident. And I am adding property loss, content loss some and, and basically that will convert everything into my data frame.

S1 **Speaker 1** ▷ 1:52:03

Group A is also useful for situations where you may want to calculate the average as well. The same thing, but instead of home, I'm gonna use me questions guys about group I pretty straightforward with loads of examples, but basically that's the whole concept. Okay? So let's do group by activity. I'm gonna open the same breakout rooms for this one. Actually we have 15 minutes to work in and then I can show you the answer. Okay? Let me create the same breakout rooms. So I'm gonna open all the breakout rooms and you know, you have the options to go to a, a room, start sharing your screen or you have the option to work individually, okay? Please, if you stay here, you can ask questions. If you go to the breakout rooms as well, ask questions to the classmates and, and please let me know if you've finished before if you want to check some, okay, it's activity. I show you that activity zero seven par census group by,

S1 **Speaker 13** ▷ 1:53:20

Could you add me to any breakout room please? No one has joined.

S1 **Speaker 1** ▷ 1:53:26

Yes,

S1 **Speaker 10** ▷ 1:53:30

Me too. Last time I got to the like break room 12, only myself.

S1 **Speaker 1** ▷ 1:53:36

Okay, in which room you are Joseph? I see, I see you here. I'm gonna send you to six.

S1 **Speaker 13** ▷ 1:53:52

Thanks.

S1 **Speaker 1** ▷ 2:10:39

Okay, we're just waiting some seconds for, we want to come back and then I can share the answer and, okay, great. Everyone is back. Let me share the answer right now. Can you see my screen thumb up? Perfect. Okay, so as you can see here, what we are doing is basically we are writing, we are reading the csv and first they ask us to create a data frame with columns to total here, count state with these columns, right? So we do a census total df, df with the filtering of this column so far you should be able to do this part without any problems. Okay? I mean this part is pretty straightforward. And then just read it. Then they ask us to do a group by, by, by here and by state. Okay? So we do here, here into the left here and state into the right, this is going to put you into the left state, into the right and then we do a zoom, right? Because we have to do a function when we using a, a group five. Okay? Now that I have that, zoom is telling me to rename the columns to make them more understandable. So I will rename population to total population and blah blah, blah. You have to do not rename parenthesis column equal and then the diction. Okay? So I'm gonna see something like this.

S1 **Speaker 1** ▷ 2:12:31

After they, they ask us to get the average of a specific columns. So the first thing that I'm gonna do is I'm gonna get those columns, okay? Here, count the state, all the, all those columns. And then I am going to apply the group by to these columns again. And I'm gonna apply I, okay, so my end data frame is gonna be this one year state. And with the mean of each of the columns that I actually select selected before, then they ask us to rename the column again. So we do exactly the same thing. Rename column, you can copy and paste the code and use this part and then just save it to a csv. Okay? Export the ference to a csv. I'm gonna export the totals and I'm gonna export the app. Questions on this part, guys? Yes. Can you

S2 **Speaker 2** ▷ 2:13:31

Scroll up to four in four?

S1 **Speaker 1** ▷ 2:13:34

Yeah, no problem. Yeah, just

S2 **Speaker 2** ▷ 2:13:36

I just want to look at the code. Okay. Okay, so you brought the columns here and can you scroll down a bit?

S1 **Speaker 1** ▷ 2:13:47

Yep.

S2 **Speaker 2** ▷ 2:13:49

Yes. Okay, got it. Thank you.

S1 **Speaker 1** ▷ 2:13:55

Perfect, no problem. Any other questions?

S1 **Speaker 10** ▷ 2:14:02

Yes, can I see the scroll? Scroll down a little bit. Eight, four. Okay. Okay. Okay. Yeah. Seven, eight. Thank you. I just need to make sure what's wrong with my call. Okay, thank you.

S1 **Speaker 1** ▷ 2:14:14

Yep, no problem. This call is gonna be hearing Gl lab so you can compare it after you taking it. Okay? Okay. Perfect guys. So let's go to the last topic of today's class that is actually sorting

S1 **Speaker 10** ▷ 2:14:32

Oh, wait a minute. Wait a second. Hugo, you say like the code will shows in where,

S1 **Speaker 1** ▷ 2:14:39

Sorry.

S1 **Speaker 10** ▷ 2:14:40

The activity that you know in

S1 **Speaker 1** ▷ 2:14:42

GitLab in GitLab are going be

S1 **Speaker 10** ▷ 2:14:44

Okay. Thank you. I just wanted to make sure. Thank you.

S1 **Speaker 1** ▷ 2:14:47

Yeah, so sorting basically is sorting like in Excel. There is nothing too much to to understand here just how you are gonna use the code. Basically sorting the data frame based on meal column will sort from lowest to highest if no other permit is passed. If you just put meals, taxes, the F equal to taxes, the F sort values by meals, it's gonna put like lowest to highest. Okay. You can change that by adding this, by sending equal false and it is gonna put it in the other way around. It's possible to sort based upon multiple columns but not really. Okay. What it's gonna be doing, basically it's gonna sort for the first one and then it is gonna try to sort for this one, you're gonna see this meals and rent count F Texas D F stored values, meals count and rent count ascending. Okay? So let's see, meals is actually okay, right? Because it's gonna be mill count and 19 and 11 and it's going down. But if you see rent count, that is not happening for rent count because is is like in in Excel, right? When you sort a column you can sort it just one column for the whole table or you can sort two columns. But these, these two columns are going to be sorted only by one column. Okay? Not by two columns. That's the only difference here. Is that clear here?

S1 **Speaker 1** ▷ 2:16:51

Yeah. Okay,

S6 **Speaker 6** ▷ 2:16:52

Well question, it'll sort by the second column. If there's like if there are duplicate meal counts, like the first two had the same meal count, then it'll sort the red count.

S1 **Speaker 1** ▷ 2:17:02

Exactly. That is correct. Yeah, it's gonna do that, but it's not going to show you 26 and then 21 because here is 26 19 and then 26 because it's leading by this valley. Okay, yeah. Perfect. So to see the sorting by multiple column better, we can compare the last data frame with the second column sort and alcohol count compared the order of 2 54 values rent count. So basically it's quite the same of what we use. So, but it is just actually checking these meal counts. An alcohol count in this case, alcohol count as you can see here is actually doing the sorting but not really, you see where is it? 52 to 54.

S1 **Speaker 1** ▷ 2:18:01

Okay. So you're gonna find those kind of values when you do this kind of sort. It's very just to sort by each one and then just create maybe a new data frame and if you want to sort by another column, you can create another column sorted by that other column parameter. Okay? So the index can be resetted really easy. You just put dot reset index drop equal two and then you, because you see the index, the index is all messed up so you can just reach the index. Any questions about sorting guys? Okay, perfect. So why don't we work in the last activity together, open this activity, search for the worst and if you open it, we can see this. You see my screen now? Perfect. Okay, so

S2 **Speaker 2** ▷ 2:19:14

No, we don't see it.

S6 **Speaker 6** ▷ 2:19:15

It disappeared. Oh god,

S2 **Speaker 2** ▷ 2:19:20

Now we see you.

S1 **Speaker 1** ▷ 2:19:23

Can you gimme three now? Yep. Okay, perfect. If you wanna open this activity, we're gonna work together. Okay,

S1 **Speaker 1** ▷ 2:19:37

So basically let me check the, the README file, this one Readme. In this activity you will take a data set on San Francisco airport with the consumption and determine which month in the dataset has the highest consumption for each utility. And they're gonna give you the values, right? Read in the CSB file provider and print into the screen, print out the list, the whole values within utility, select a value from the list and create a new data frame that only includes the, that utility. Note that co utilities have more than one option for owners. So you should limit this new data frame to single owners such as tenant for the data frame based on the LE level of consumption from most to list. So this is the sorting, reset the index and print out the details. Okay, let's work on this.

You ready? So who can help me explaining me this code please? This piece of code

S1 **Speaker 14** ▷ 2:20:53

That's telling her where to go to find the file, correct?

S1 **Speaker 1** ▷ 2:20:57

Yeah,

S1 **Speaker 14** ▷ 2:20:58

To read.

S1 **Speaker 1** ▷ 2:21:00

Okay. What is this variable doing?

S1 **Speaker 14** ▷ 2:21:06

Converting it. So, and it can read it?

S1 **Speaker 1** ▷ 2:21:10

Yes. And it's converting this into a data frame. Okay. One important concept you have to understand is that this line is converting my CSV into a data frame, okay? Because I am using TDA here. Okay? That's a key part because sometimes you track this data frame to your code and you don't know that you are working with the data frame. But yes, here you are no longer working with E L B, you're working with the data frame and yes, it's different guys. Even though it looks the same, it's different. Okay, thank you very much for that Adrian. Then the next part, collect the list of all the unique values in utility. What this doing,

S4 **Speaker 4** ▷ 2:22:04

You're getting the unique values in the utility column and you are giving that as an output of as a list.

S1 **Speaker 1** ▷ 2:22:12

Perfect. Yeah, yeah. Yes, exactly. You are actually giving this as a list. Why? Because you are going to use it. Okay, so this list, basically it's gonna contain the data frame that you just create. You are going to select the column that you want to work in or you want to do something to that column and what, what function you're gonna u do the unique function, okay? And then you get passengers, gas, electricity, and water. That's it. Okay? Now we want to put these utilities into a frame. So this, this is one thing that we haven't before. So the only thing that we have to do is put the consumption D f utility. I am getting to that because here is a list, right? I haven't put this into a, I haven't put this into a variable, into a data frame, anything like that. So I am again, taking only utilities column, I am getting the value count and now I am passing that into a frame but not storing this into a variable. This is important guys. This, I just can see this part in the output, the frame, but it is not safe anywhere. Okay?

S1 **Speaker 1** ▷ 2:23:46

So far, so good. Any questions for the moment here?

S1 **Speaker 12** ▷ 2:23:53

Okay, is this, Hey Hugo, sorry, with the two frame that you've done, can this be done by doing PD data frame?

S1 **Speaker 1** ▷ 2:24:02

Totally, but you are going to save that PD data frame into a bar, okay?

S1 **Speaker 12** ▷ 2:24:09

Oh, so this, this,

S1 **Speaker 1** ▷ 2:24:11

This is not saving. Yes, this is not saving anything to a variable, it's just showing the data.

S1 **Speaker 12** ▷ 2:24:17

Ok, thank you.

S1 **Speaker 1** ▷ 2:24:20

So looking only at electricity consumption with an an owner here guys, what I'm gonna do is consumption D F using the same data frame. You see I am not using any different things. Consumption d f lock. And now I am going to access through headers. That way I'm using lock consumption D equal to electricity using my operator consumption Bs owner equal. Okay? And it's gonna show me electricity and owner at. So good so far. So far so good. So now the consumption, eh, this is exactly the same thing. Exactly the same thing. They just repeated it for some reason it's a bar. Remember about using the equal. Equal, okay? Because the equal, equal is important in Python.

S1 **Speaker 1** ▷ 2:25:30

And now what I'm going to do is consumption d D usage. I am going to divide it by a hundred so I can read it better by, so I can read it better because it's in very high units. So consumption df, this is an example of how you can just do a simple operation to a whole bunch of data that belongs into a series consumption, DF usage slash 10,000. And basically you're gonna get it like in a much presentable way. And now that you get it in a much presentable way, you can sort it, okay? Sort these values usage in the column attending full. Now you'll see that I am storing this into a new data frame, data frame number two, I'm going to the index data frame. Index drop through and voila, I have my table an I just have this into questions about this piece, this block, and this block. Questions about those?

S1 **Speaker 15** ▷ 2:26:47

I have a question.

S1 **Speaker 1** ▷ 2:26:48

Yes.

S1 **Speaker 15** ▷ 2:26:49

So since you divided the usage by 10,000, does that mean you have to change the units?

S1 **Speaker 1** ▷ 2:26:57

Yes. Yes. That's a good question. And yes, I think you have to change the unit that we didn't do. Yeah. Alright, good spot. Any other questions?

S1 **Speaker 1** ▷ 2:27:21

Okay, perfect guys, well you can go and redo this activity by yourself. We are gonna put all the solutions already and for today's class, that's it. Actually we're gonna end early today. That's great. Great work there guys. Again, I am going to open the four breakout rooms with 40 days in it so you can go to each of the rooms if you still have questions. If you don't have questions, please feel free to go and see you. Until next Tuesday guys, thank you very much for today's class. Have a great, great night. Have a great long weekend. Rest. Go through the activities again if you can. If you don't, that's okay, but see you on Tuesday guys. Thank you very much. Have a great night.

S1 **Speaker 11** ▷ 2:28:16

Good. Will I be able to pull all the the solutions today? And

S1 **Speaker 1** ▷ 2:28:22

I am, yes, I think so.