Python and AI Power-Up Program Offline Class-20250829_113158-Meeting Recording

August 29, 2025, 6:02AM

2h 11m 9s

◉ **Tirth** started transcription

Tirth 0:05

OK, yeah, let's get started.

Tarun Jain 0:07

OK, so I'll just brief out what we will be doing today. So we'll start off with the quiz.

So did anyone try out some of the new commands from pandas or matplotlib?

I mean Nampai.

Tirth 0:22

OK.

Mitesh Rathod 0:23

Um, not get a chance. Sorry.

Tirth 0:25

No, yesterday when it was cancelled, we didn't get the chance at all then.

Tarun Jain 0:25

OK, So what will you guys do?

Got it. So what we'll do is we'll look at the existing pandas and umpire commands and then today we'll look at some of the important filtering techniques in pandas and then how to plot certain graphs. So once we are completed with these two, today we'll also start off with EDA.

Tirth 0:34

Mhm.

Mhm.

Tarun Jain 0:47

So EDA is very important. As I said, EDA stands for Exploratory Data Analysis. Whatever you do in Excel, right? Filtering, trying to understand what kind of columns is there. So we will play with data, try to understand what columns are required. So sometimes what happens is let's suppose you have two columns.

Which are important. Out of these two columns you can create a new column, right? So this is called feature engineering. So all this happens during EDA, whether it is data pre-processing or data enhancement, that is what EDA is about.

And once we take one example today, we'll also have a second assignment. So over the weekend probably you can pick one of the data, which is very fun data set to be honest. And then what you can do is you can try to understand what the data is about, perform certain filtering technique.

And that will be the second assignment.

Mitesh Rathod 1:40

OK.

Tarun Jain 1:41

OK, so let's get started with the quiz. So can anyone tell me what will be the output of this? So how many elements do we have here?

Tirth 1:51

50 element.

Mitesh Rathod 1:52

OK.

Tarun Jain 1:52

Zero to 49, which is total 50 elements and if I want to print the number of elements, what is the comment?

Mitesh Rathod 1:54

Good.

Tirth 2:03

OK.

Tarun Jain 2:04

A red outside. OK, so the answer for this.

Mitesh Rathod 2:11

Bytes.

And.

Tarun Jain 2:19

What is the size of one element?

Tirth 2:23

Is it 8 Windows 16 bit bytes?

Tarun Jain 2:25

Yeah, it's 8 bits.

Tirth 2:27

8 bit length.

That is one byte, so it will be 15.

Tarun Jain 2:33

15 to.

Tirth 2:35

I will love 4400, so 400 bytes and 400 bytes. I don't, yeah.

Tarun Jain 2:40

Yeah, it will be 400. Why? Because it is 15 to 8.

Tirth 2:44

Alright.

Tarun Jain 2:45

So how do we calculate that? You have an item size. Every single element that you have has 8 bytes, so you just have to multiply it with array dot of size.

Mitesh Rathod 2:48

It.

OK.

Tirth 2:51

Mm.

Um.

Tarun Jain 2:57

OK, so now what kind of dimension is this?

Mitesh Rathod 3:01

three-dimensional.

Tarun Jain 3:05

So how do I print it?

Mitesh Rathod 3:08

Print dot shape.

Tirth 3:10

Array 3D dot.

Tarun Jain 3:10

No.

Mitesh Rathod 3:11

Yeah.

OK.

Tirth 3:15

We need dimensions, no? We want dimensions and.

Tarun Jain 3:15

So what will shape do?

Hi it should be NDM 3D of NDM.

Tirth 3:23

Yeah.

Tarun Jain 3:25

So this will print 3. Now what you need to do is you need to guess what is the shape of this particular array.

Tirth 3:33

E.

3 cross 3 cross 5.

Tarun Jain 3:43

How do we predict it? First you have to count all the 2D arrays you have within the given third bracket. So we have total 3. So it is 3. Then what you need to do is you need to check every individual 2D array and what is the shape of it. So totally we have one.

Tirth 3:48

Right.

Mitesh Rathod 3:48

Yeah.

Tirth 3:49

I like.

Tarun Jain 4:01

2-3 rows.

And how many columns do we have total 5?

Mitesh Rathod 4:03

Let's see.

Goodbye.

Tarun Jain 4:10

OK, so this is last in numpy.

So what is the shape of this?

Tirth 4:34

It is 1 cross 50.

Mitesh Rathod 4:35

OK.

Tarun Jain 4:36

It's just 50.

50 comma and this is 1 dimension. What about this one?

Tirth 4:42

M.

The 25 rows and two columns, so 25X to 25 comma 2.

Tarun Jain 4:52

Uh, 25 comma 2.

And this is fluids.

Tirth 4:56

OK.

Tarun Jain 4:59

And what other combinations can I try here apart from 25 comma 2?

Tirth 5:04

5 comma 10.

Tarun Jain 5:07

So this one everyone are aware, right? So whenever you create any shape, you just have to ensure the reshape also has to match the number of elements.

Mitesh Rathod 5:10

OK.

Tirth 5:10

Yes.

Mitesh Rathod 5:15

Right.

Tarun Jain 5:20

OK, so if I have double equals, what will be the data type of this entire line?

Mitesh Rathod 5:30

Data that it'll be boolean true.

Tarun Jain 5:35

Yes, it will be boolean.

So now what is the final answer?

Mitesh Rathod 5:42

For us and by.

Tirth 5:44

And N bytes remains the same, so it'll be true.

Mitesh Rathod 5:47

It will be the same.

Tarun Jain 5:48

Correct. So N bytes, as I said, if you look at reshape right, how much size does it have?

I didn't plan this, but let me print this outside.

Mitesh Rathod 5:59

Yeah.

Tarun Jain 6:05

The size still remains the same and for any array.

Mitesh Rathod 6:05

15.

Tarun Jain 6:11

The item size is always 8.

Tirth 6:12

Hmm.

Tarun Jain 6:13

So 15 to 8, 400. So both array and reshaped will have N bytes to be 400.

Mitesh Rathod 6:21

So.

Tarun Jain 6:22

OK, so let's quickly look at some of the Pandas command as well.

So how many columns do I have and how many OK rows? It will be very difficult. So how many columns do I do I have?

Tirth 6:36

3.

Mitesh Rathod 6:37

Yeah.

Tarun Jain 6:38

It's just three. So I have anime name total episode genre. So total we have total 3 columns. Now if I want to make total episode as index, what function do I have?

Tirth 6:39

M.

Hmm.

That index.

Yeah, yeah.

Tarun Jain 7:01

Is this clear?

So whenever you think you have a data set which has customer ID, serial ID and you don't need the default index, then what you can do is you can directly use DF dot set index and the particular column name.

Mitesh Rathod 7:20

Thanks.

Tarun Jain 7:20

Now someone else apart from Teer should answer this. What is the output of this?

Mitesh Rathod 7:26

It will be apply for uh uh.

For to all groups, no.

Tarun Jain 7:32

Cross check one more time.

Mitesh Rathod 7:34

No.

Greetings. It will apply the area of four all the groups.

Tarun Jain 7:37

What is the variable?

Can you repeat?

Mitesh Rathod 7:44

Uh, it will apply area of four to all the groups, I mean least.

Tirth 7:49

Uh, can I answer now?

Tarun Jain 7:51

No, not it.

Will this be error or will it be over?

Tirth 7:56

It will be a good. It will be a good.

Mitesh Rathod 8:00

7.

Tirth 8:00

Edit.

Mitesh Rathod 8:02

OK, same column size not.

Tarun Jain 8:02

So one thing we have to notice, let's suppose I have total I'll do DF dot shape.

Mitesh Rathod 8:09

OK, OK.

Tarun Jain 8:10

So the shape is 18 rows and three columns. OK, now when I do DF dot rating and if I give a sequence right? So now what DF will expect? Hey, you're giving sequence and total I expect 18, but you're giving only one.

Tirth 8:12

It.

Mitesh Rathod 8:14

Columns.

Tarun Jain 8:27

So it's an error.

Tirth 8:28

Mm.

Mitesh Rathod 8:29

Hmm.

Tarun Jain 8:29

But if I give a static variable, now what will be the output?

Mitesh Rathod 8:32

It will. It will be applied to all and default values of Japan.

Tarun Jain 8:37

So it will create a new column first of all.

Mitesh Rathod 8:40

For parties.

Tarun Jain 8:42

And then it will add Japan for every single.

Entry. Is this clear?

Mitesh Rathod 8:46

Both.

Tirth 8:48

Yes.

Tarun Jain 8:49

So if you're creating a new column, you have to ensure you match the sequence, because most of the time you will not use static variable, you will have the actual values.

Right. So it's always better to have sequence with same number of size right length of the DF.

Mitesh Rathod 9:07

Right.

Tarun Jain 9:10

Now what will be the output of this?

Mitesh Rathod 9:14

It will print only the uh enemy in January, but it will be in no.

Tarun Jain 9:21

This is an error.

Why? Because if I want to print 2 columns, I need to give it insider.

Mitesh Rathod 9:32

Double packets, OK.

Tarun Jain 9:33

Double bracket. So if I'm just printing one of them then I need just.

Tirth 9:34

List, yeah.

Tarun Jain 9:40

Single bracket, but if I'm using two.

Then it it has to be inside double bracket.

When is this used? When you are using 2 columns and then you're creating a new column from these two columns.

Is this clear?

Mitesh Rathod 9:58

Yes.

Tirth 9:58

Yeah.

Tarun Jain 9:59

OK, so we'll quickly move on with the the same data that we had. I mean, not data, the collab.

So we actually left it here right at series.

So what we'll do is we'll start from reading CSV file.

I'll share this repo URL again if in case you need to download the data.

So in this particular URL you have total 3 CSE files. One is uh data dot CSE, then you have random dot CSE and then you have narutoanalysis dot CSE. So what you can do is you can download all three of them.

And once you download you have to click on this directory icon.

And then upload that particular data.

Let me know once it's done.

So far whatever we have done, we were just playing around with data frame. But if you have structured data, why do we use pandas? Whenever you are playing with structured data like CSV file, Excel or if you have any XML or Jason file and if you want to display that in a tabular format just like your Excel.

Mitesh Rathod 12:03

OK.

Tarun Jain 12:16

During the time use data frame right and we have different ways to read it as well like pandas dot read CSV read excel.

Mitesh Rathod 12:24

It's what the one.

Um.

We need everything.

No.

Mm-hmm. One one primary.

I mean.

Enjoy this.

Tarun Jain 13:34

Is it done?

Mitesh Rathod 13:36

Just a second.

At least.

Right.

Yeah.

And what?

It was.

RamKrishna Bhatt 14:07

Approved all three it outside on the right so.

Tarun Jain 14:12

What happened?

RamKrishna Bhatt 14:12

Is that fine? I just have oh, I just uploaded all three at the root level. Is that correct?

Tarun Jain 14:19

Yeah, here on the right all the three files. Yeah, yeah, that works. So what you can do is you can create a new cell.

RamKrishna Bhatt 14:20

Yeah, yeah.

OK.

OK.

Tarun Jain 14:28

And then just type a list.

Mitesh Rathod 14:28

OK.

Tarun Jain 14:31

You should see those files.

Mitesh Rathod 14:31

Bye.

RamKrishna Bhatt 14:35

OK.

Mitesh Rathod 14:36

Yeah.

Tarun Jain 14:36

If you're able to see it, you can read it directly.

Mitesh Rathod 14:47

OK.

Tarun Jain 14:49

Whenever we are running any command in VS code, we just have to make sure that we have to start it with external meter mark.

Like LS is 1 and if in case you want to check the current working directory which is PWD, it is content.

So if you want to.

Mitesh Rathod 15:07

We need to upload all three CSV file, right?

Tarun Jain 15:14

Uh, you can upload all three.

Mitesh Rathod 15:16

OK.

Tarun Jain 15:20

They won't let me upload it.

Mitesh Rathod 15:20

Sorry.

Please.

Tarun Jain 15:27

I'm

OK.

Mitesh Rathod 15:36

It's a 77.

OK.

Tarun Jain 15:48
Should we proceed?

Mitesh Rathod 15:49
You know.

Tarun Jain 15:53
Uh, is it done or anyone is still uploading?

Mitesh Rathod 15:55
Yes, sure.

Tirth 15:56
We can proceed.

Tarun Jain 15:57
OK, so now if you look at the particular directory that we have, we have total 3 files. One is data dot CSV and Naruto analysis CSV and then random dot CSV. If you want to display this right, what we can directly do is first we need to import pandas.
Import pandas.
As PD and I hope everyone has this collab notebook right? We are starting from here reading CSV file.

Tirth 16:25
Yes.

Mitesh Rathod 16:26
Yes.

Tirth 16:28
Yeah.

Tarun Jain 16:31
And then once you import pandas as PD, we just have to use a function called read

CSV. There are two ways to do it. Whatever commands we have in data frame that is also supported in data. If not, you will see some of the notebooks where people will convert this into.

Mitesh Rathod 16:38
Mhm.

Tarun Jain 16:50
Data frame.
OK, both the ways are possible, but we don't have to create an additional variable to do all the operations, right? Because even data has the same functionality that a data frame provides. Is this clear?

Tirth 17:09
Yeah.

Tarun Jain 17:10
OK, so this again is very simple syntax which we have already covered. We have data dot head which will display starting by head and we have total 2.

Mitesh Rathod 17:18
How you?
Is the number.

Tarun Jain 17:24
Columns right? And just like that we also can increase the number of head if we need. Here I'm giving 10. It will display 10. We have already covered head as well. Tail is also covered. It will display the bottom five and if you want to display any other number, you can also provide that number inside the brackets.

Tirth 17:25
Mm.

Mitesh Rathod 17:25
OK.

Tarun Jain 17:44

So red was done, tail was done and we also add sample which will randomly select only one by default. But if you want to generate multiple samples you can provide any integer number.

Tirth 17:55

M.

Tarun Jain 17:57

Till here anyone has any doubts? We just read this CSV file and we are displaying it and there are three ways to display it, head, tail and sample.

Mitesh Rathod 18:05

Mhm.

Tirth 18:06

But.

Mitesh Rathod 18:07

Mhm.

Tarun Jain 18:08

OK, so now whenever it comes to understanding DDA, the first thing is we have to check the data type of it, data type of each column.
And then what is the description?
Of the given data.
So this is where we have two new function. So these two functions that we have right? Usually how we proceed with this is first step is.

Mitesh Rathod 18:35

OK.

Tarun Jain 18:44

You read the data.

Second step is.

You copy the data.

So for example, you just have to do copy DF equals to data dot copy.

And then copy DF is also same.

Sample.

Mitesh Rathod 19:10

M.

Tarun Jain 19:12

It's discipline, right? So the second step is.

Mitesh Rathod 19:14

Mhm.

Tarun Jain 19:17

Copy your data. Third is display. It can be.

Head or tail or sample. Then you have to check the info and the description which is described just to get some understanding of OK, what is the mean of certain column? What is the median of certain column? What are the different data types I have?

So if you have any time right, as I said, whenever it comes to pandas, some of the data set is time based, right? So during time based you'll have to tackle it separately. I'll show you how to tackle that. So all these data types can be.

Reported using these two functions and then you check for if there is any empty data or not.

If you have your empty data, then you tackle how to tackle.

It.

The empty data.

And then you proceed with EDA. EDA, there is no what you call. There is no procedure. You can start from anywhere. You can start with visualization. You can start with feature engineering where you use existing columns and you create a new columns. Tiller, is it clear?

We are just taking the steps on how we usually have to proceed. So now if you see if

I print data dot report it is displaying arc name then total episode. So we can also have the movie data. What you can do is you can do movie data.

Mitesh Rathod 20:34
Yes.

Tarun Jain 20:50
So what is the comment? I want to read this data file.

Mitesh Rathod 20:56
Let's see.

Tirth 20:56
TD dot read CSV.

Tarun Jain 20:57
dot read CSV data dot CSV.

Tirth 20:59
It.

Tarun Jain 21:03
And now if I just print movie data dot info, the reason why I'm reading this now is there are more columns in data dot CSV file. So if you see you have show ID, then you have type, then you have title, you have director, then you have cache, you have country.

Then you have date added, then also you have release date. So there are so many columns that you have. So one interesting thing is when you display this, for example I'll display sample to be true.

So date added. What is the data type of this?

What was the data type?

Movie data.

dot info.

What was the data type of data added? It's object. So when we mention object, it is nothing but string. Now specifically this is date, so this has to be converted into a date format which we will do in the coming comments.

But you understood when it comes to info it will just tell you OK total you have 8800 non null. That means the total shapes that you have right? It's around 8800 but if you look at director it shows.

Mitesh Rathod 22:10

OK.

Tirth 22:11

Mhm.

Tarun Jain 22:25

6000 only and then you have cached which is 7800. That means this three columns this four as null data.

And there is also rating and duration which has four empty and here there is 3 empty.

So you understood why we use info?

Tirth 22:45

Yeah.

Mitesh Rathod 22:46

Mm.

Tarun Jain 22:47

Next one is describe. So describe it's mainly when you use outliers, right? Outliers in the sense if you notice very uneven distribution and we took the example of Titanic last time and in Titanic I showed you what will be the median. So 50% in the sense this.

Is median 25% is nothing but quartile. 75% is nothing but second quartile. Maximum is nothing but let's suppose.

I have data.

And here you have episode, right? Total episodes.

So the maximum number of total episode will be 85 is what this Max value is about and when you mentioned describe if you see it is only taking numerical columns.

Mitesh Rathod 23:39

Mhm.

Tarun Jain 23:41

Uh, you saw this difference?

Mitesh Rathod 23:42

Mhm.

Tarun Jain 23:43

So when I do describe even though my data has two columns, so if I just print data I have total 2 columns but still describe will only generate the report for numerical column which is your integer.

Why? Because you can't find out median first quartile, second quartile for a categorical column. So this thing categorical in the sense string.

Now if I do the maximum of it, let's suppose I want to print total episode dot Max. It is 85. This is what it is doing here and then median is nothing but 18. Minimum is nothing but zero is also there and this is standard deviation and this is mean. So this too will tell you some information about the distribution.

Mitesh Rathod 24:34

Yes.

Tarun Jain 24:43

So mean and medium should always be very close. It should not make much of the difference. If there is much difference, that means your data has outliers.

Which is below the threshold or above the threshold values.

Till here everyone is clear info and uh describe.

Tirth 25:04

Yeah.

Tarun Jain 25:06

Here I've already imported it data set equals to PDCSV data and we are repeating the same step.

Mitesh Rathod 25:08
OK.
Yeah.

Tarun Jain 25:14
If you see data salt describe works only on numeric data. So how many data I have here? I have total 11 columns. Out of 11 columns there is only one column which is in numerical.

Mitesh Rathod 25:26
Yes.
Who's listening?

Tarun Jain 25:29
Right, so now if I do this, I'm getting data set describe only for release here.

Mitesh Rathod 25:30
Yes.
Hmm.

Tarun Jain 25:39
Till here everyone are following.

Mitesh Rathod 25:42
Yeah.

Tarun Jain 25:43
OK, so this command also we have checked earlier. If you want to print the shape of it, you can directly do data dot shape and then it will display the number of rows you have and the number of columns all the data frame.
And your CSV file, CSV or Excel will always be.
In 2D, which is only rows and columns. There is no 3D when it comes to structured

data. Why? Because it is tabular column.
Tabular based data set.
You will never encounter.
3D data frame.
I'm not sure because I've been working with pandas for probably three to four years. I've never seen any 3D data from till now. Most of the time you get a CSV file. CSV file is just an Excel sheet. So what do you see in Excel? You only see rows and columns in Excel.
Alright, so now as per whatever you mentioned earlier, let's suppose if I come back. What was the first step? Uh, I'll just write procedure.
We start with first step, then second step is copy the data, third is display, then we have info on describe and 3rd is to check the empty data or not and now what we can do is.
Or does anyone remember the comment to check the empty data?

Tirth 27:26
Yes it is is null and some with some so.

Mitesh Rathod 27:26
Please.

TJ **Tarun Jain** 27:28
Gita is Nal Andan Dodsam.

Tirth 27:31
Sun.

Mitesh Rathod 27:33
Some.

TJ **Tarun Jain** 27:35
So the arc names is 0, then the total episode is also zero and now if I do data set.
So show ID 0, type is 0, title is 0, director has around 2600, cache is around 800 and here we run one more test. Let's suppose you're building a recommendation system. OK. Or you're building sentiment analysis?

So for sentiment analysis you have tweet tweets dot CSV file and inside this you have text and you have other important columns if you encounter any column which has more than 20%.

Tirth 28:11
Oh.

Tarun Jain 28:21
Of empty data, not 20, I'll make it 30. So if it has more than 30% of empty data, that means you can remove that column because that will not add any meaning to it, right? So.

Tirth 28:35
Mm.

Tarun Jain 28:37
This is where null is usually preferred and we will do this in EDA as well. There will be times when you will have some of the data which will have around 60 to 70% empty and that will just add no meaning for any model building. So what you usually do is. If there is any logic to fill the data, you can use fill in there.
Then there is also a technique called Imputer. So these are the two techniques. You have certain strategy if you know OK, this column has 60% empty data, but in order to build it, since you're a domain expert, you will know this column is very important. And if you know the logic to fill the empty data, then you can keep that column and in order to fill it we have fill in A and imputer which we will cover.
This probably will not cover, but fill in a is something we will cover which we have already covered earlier as well.

Tirth 29:30
We did cover it. Yeah, we already covered it.

Tarun Jain 29:36
So if you just print is null you will get false, false, false, true, true. But what I need is I need the actual sum.

Mitesh Rathod 29:40

But.

Tarun Jain 29:49

OK, till here is it done.

So now what we will do is we'll try to filter out the data. So what do I mean by filtering? Let's suppose you have sales data.

And you need to know.

How many entries?

You have that is in India.

And.

First of all, let's write the columns you have. You have sales ID.

Then you have customer ID.

Then you have country.

Then you have payment or I'll just tell invoice.

And you have sales and then you have some important columns. So here what typically you do in filtering is you want to know how many entries which is your customer ID.

Are based out of country India and.

Are paying invoice.

More than.

10,000 or 5000 right? So here if you see how are you applying the filter, you have a data frame. You have to check the country first if in case that country belongs to India or not.

And then invoice should be.

More than 5000 and then you just have to display those particular, uh what you call display the particular data frame and then you can check the length of it.

Is this clear?

Tirth 31:28

So it is this. Is this set separated by space? That's it.

31:31

Yeah.

Tarun Jain 31:32

Welcome to the logic you'll have and operator.

Tirth 31:35

OK, OK.

Mitesh Rathod 31:35

OK.

Tarun Jain 31:37

So this is just example I was giving. This is not actual syntax.

Tirth 31:38

OK, OK, OK.

Tarun Jain 31:41

OK, so the best example is you can. Are you guys using Spotify?

Tirth 31:47

Yes.

Tarun Jain 31:48

So you have this Spotify app, right? So in Spotify app, how is it able to get your top five artists, then the total number of hours you're spending. So these are some of the filtering technique. Let's suppose you have Spotify data.

Mitesh Rathod 31:49

Yeah.

Tarun Jain 32:05

And now for the given person, let's suppose Tarun, I want to know who are the top five artists, then what are the total number of music he has heard. So once you have the data, you can do the filtering, you can try to get the actual amount of count. This is very similar to your SQL commands.

Tirth 32:25

Mm.

Tarun Jain 32:26

Right when you use where you use select, you give a condition and then you define where right. This is similar to that and then you can just print the length of it to get the actual amount of some kind of number right for verification.

Mitesh Rathod 32:39

M.

Tirth 32:41

Mm.

Tarun Jain 32:42

This is similar to SQL command and this is very important when it comes to pandas.

Tirth 32:43

Mhm.

Tarun Jain 32:48

And what we will do is we will check two of them. One how to apply filter or masking for only one single column and how to define it for multiple columns. One second.

Tirth 33:01

OK.

Data frame.

Mitesh Rathod 33:26

When I was.

Tirth 33:35

So what's going on by link covers?

Big by X equal to 10.

Mitesh Rathod 33:48

2 minutes. OK, OK.

Dubai.

RamKrishna Bhatt 34:14

They're both famous, bilingual.

Mhm.

Tirth 34:24

Big Bay.

Mitesh Rathod 34:29

And it's a coding Sanskrit.

Unicorn.

RamKrishna Bhatt 34:49

Cool my subject Mamuk sentence.

Tirth 34:57

Hello.

Yes, yes.

Mitesh Rathod 35:00

Hello.

TJ **Tarun Jain** 35:00

OK, so here I'm again using the same data that I use for quiz. So I have total 3 columns. So here what I'll try to do is I'll filter like OK which anime belongs to Adventure then.

How many animates has more than X number of total episode and then I just have to filter it out.

So this is just basic example. We will also filter it out based on the actual video game sales data set as well when we perform EDA, but this is just to get the syntax familiarity.

So this is single filtering. You can either call it filtering or what is the best way to describe this?

Filtering or grouping? What do you do this in SQL?

Mitesh Rathod 35:48

What is?

Tirth 35:50

Filtering if it is.

And it is filtering you. Yeah, it is filtering.

Tarun Jain 35:58

OK.

So now what we have is we have total 17 entries and you want to know in the given data that you have right, the data will obviously be more than 5010 thousand entries.

How many anime episodes have more than underrated episodes?

Mitesh Rathod 36:04

Huh.

Tirth 36:07

Mhm.

Mhm.

Hmm.

Tarun Jain 36:17

So what is the condition here? You start with DF. So the first thing one needs to understand if I'm grouping two different columns, let's suppose I have.

Tirth 36:30

And my name is Johnner.

Tarun Jain 36:31

Huh. Any of my names and genre? What was the syntax?

Tirth 36:36
A comma.

Tarun Jain 36:39
So if you see here, these are the two important columns, but if I want to display it from the data frame, I start with DF then square bracket.

Tirth 36:49
Mm.

Tarun Jain 36:49
So this is the syntax to display, which is simple indexing.

Tirth 36:54
Oh.

Mitesh Rathod 36:55
OK.

Tarun Jain 36:58
Right so here also first what you have to do is just define your data frame and square bracket then add your condition. So the condition is I have a column called total episodes and I want to only display. So this thing is.
Only display which is true for this condition, right? If it is false, just remove it.

Tirth 37:17
Mm.

Tarun Jain 37:21
So if I print this out separately, it is similar to what you saw in is null. It will just add true, false, true, false.
So what I need to do is only display those which are true.

Tirth 37:33

Um.

Tarun Jain 37:35

So now you can only display those anime names which has more than 100 episodes and then you can do length of this thing.

And then tell hey our platform has total 5 anime which are more than 100 episodes.

Probably this will be more, but in our data set it is just 17 entries.

Tirth 37:53

Mm.

Mitesh Rathod 37:54

Mhm.

Tarun Jain 37:54

And there is second way to do it, which we saw in function as well, which is your map function and Lambda. So what is the syntax of Lambda? You have Lambda, then you have a variable. So this variable has to match with the data that you are using. Let's suppose you have.

Total episodes.

So now every single entry you're labeling it as X. OK, so this is 1 entry X now and then you're incrementing it.

Tirth 38:32

Mm.

Tarun Jain 38:33

And then the condition is colon. I need more than 100 and in order to apply this Lambda you have a function called map.

Mitesh Rathod 38:38

M.

Tirth 38:46

Hmm.

Tarun Jain 38:46

And now if I print this again, it's true, false, true, false.

Mitesh Rathod 38:50

Need to grab into PS.

Tarun Jain 38:53

And then you have to wrap it inside DF. Is this clear?

Mitesh Rathod 38:57

But.

Yes.

Tarun Jain 38:59

So what change did you notice?

Here.

OK, it's same.

Tirth 39:11

Oh.

Tarun Jain 39:15

And now what we have to do is uh.

We have to apply multi filtering. What I need to try to do is I want to display which are more than 100 and then which belongs to Shonen. This is the second condition and then I only have to display those.

Tirth 39:31

Mhm.

Tarun Jain 39:31

There is also one more logic which is very tedious and probably not recommended.

So what is series?

Can anyone recall what is series?

Mitesh Rathod 39:42

Oh, no.

Tirth 39:51

No, OK.

Mitesh Rathod 39:52

No.

Tarun Jain 39:53

OK, so you have data prim.

So data frame usually handles multiple columns.

At once, whereas series is just.

One individual.

Follow.

Uh, where is that examples?

6.

So if you see it, you have a series called good, bad and neutral and if you.

Mitesh Rathod 40:28

Mhm.

Tarun Jain 40:34

In 14 um.

So if I display DF, now what is DF? It's good, bad and neutral. So the data type of this is series. There is only one single column.

Mitesh Rathod 40:53

Yeah.

Tarun Jain 40:54
Is this clear? Series is only one column, data frame is multiple columns.

Tirth 41:00
M.

Tarun Jain 41:10
So the first approach was this is the easiest approach, just define DF and then the condition that you need and the second approach was the map. So map is mainly used when you will have multiple columns to filter out that during that time map is very useful.
Function right? Whenever you have to do data engineering, let's suppose you have three columns. These three columns are very useful and it can create a new column which is even more useful. So during that time again you'll use map and series.
This approach is not recommended, So what you're trying to do here is you're manually looping it. So what you're trying to do here for total episode and then you're checking a condition. If it is more than 100, then append true. If it is not more than 100, append false.
So this is a simple list. This is this has nothing to do with data frame if you notice. So you have this particular column. Now if I just print this, what is the output of above 100 check?

Mitesh Rathod 42:14
Empty list.

Tarun Jain 42:16
It's an empty list and what we are trying to do is I want to loop through entire episodes. If it is more than 100 then up and true, if not up and false.
And now if I run this, you have true false, true false.

Mitesh Rathod 42:28
But it.

Tarun Jain 42:29

And then you just have to pass it over series. So now we are creating a series. So this series has true, false, true, false. And then once you have those true false condition, you are only adding it in your data frame and once you have it in data frame you can display the entire column.

Mitesh Rathod 42:42

Yeah.

2.

Tarun Jain 42:48

This is manual effort. If in case probably this is not preferred as well because map will handle most of the excuses.

Mitesh Rathod 42:56

It.

Tarun Jain 42:57

For single columns.

Tirth 43:00

The map can be used for multiple columns as well.

Tarun Jain 43:03

Yeah, for multiple columns as well. So all these three approaches that you have right, it can be used for multi multi column filtering as well.

Tirth 43:04

OK.

OK, OK.

Tarun Jain 43:11

But this one is not recommended.

Tirth 43:14

OK.

Tarun Jain 43:15

But at the end of the day, the approach is same.

Here you'll have to add multiple lines of code.

OK, so now if you see here, I'm trying to display all the total episodes which are more more than 100. Then I have one more condition which is and and then DF should be shown in. So what kind of operator is this?

Mitesh Rathod 43:39

OK.

End up later.

Tarun Jain 43:44

Uh, what?

Tirth 43:45

But bitwise operator?

Tarun Jain 43:47

This is bitwise operator.

Mitesh Rathod 43:51

Mhm.

Yeah, we're good.

Tarun Jain 43:55

This is also bitwise in Python and or is how you usually define a logical operator.

Tirth 43:58

Good.

Mitesh Rathod 44:02

There is no.

Tarun Jain 44:19

Is this clear? And now if I print it will only display 2 columns.

Tirth 44:25

We can also replace end with end and end or not.

Tarun Jain 44:28

No, that will throw error.

Tirth 44:29

OK.

OK.

Mitesh Rathod 44:36

Same ambiguous use empty bool.

Tarun Jain 44:40

A what?

Mitesh Rathod 44:41

I'm just checking the error message.

Tarun Jain 44:45

OK.

Mitesh Rathod 44:50

OK.

Tarun Jain 44:51

So you can't apply and between 2 series.

Tirth 44:55

Yeah.

Mitesh Rathod 44:58

Use a dot MT Bolin nitrans and Li.

Tarun Jain 45:04

And in order to use map and Lambda, what we have to do is first I'm checking with data frame which is total episode. This syntax is same. This is same that you're checking for only 50. I mean you're checking with total episode that is less than 50. Here you're trying to check the short enemies and the genre should be drama. So what is the syntax? You define any variable and once you define that variable, what is the condition of that variable? Here if you see here also you have a variable.

So this variable can be anything. So here I can also write AI can define a double equals to drama. So this is looping through every single entry and then checking whether that particular entry has more than 50 episodes.

Mitesh Rathod 45:38

Yes.

Tirth 45:48

Mhm.

Tarun Jain 45:54

And it also have also it belongs to drum or not.

So now if you see you only have two animes that has less than 50 episode and belongs to genre animal.

Tirth 46:00

OK.

OK, I also sent one in the chat. Just want to know like would this work as well? Like would you have multiple variables for the Lambda in this scenario?

Tarun Jain 46:21

What is the error Lambda?

Mitesh Rathod 46:26

Uh, I get under couple, I guess.

Tirth 46:26

So.

OK, so Lambda then tuple X comma Y.

Tarun Jain 46:32

No, Lambda is a function, so it should start with function itself X comma Y.

Tirth 46:37

X comma by should be a double.

Mitesh Rathod 46:38

Uh, it might work.

Tarun Jain 46:44

No Lambda. Here you'll just have variable.

Tirth 46:50

We just keep it.

Tarun Jain 46:51

X is OK X is this one which is total episode.

Tirth 46:57

I'm trying to. I'm trying to see if this kind of syntax would work as well, no.

Tarun Jain 47:04

OK, wait X. It's considered genre.

Tirth 47:10

Yeah.

Mitesh Rathod 47:21

Par this list can can we do under square brackets? I guess it is list arguments.

Tarun Jain 47:31

No, I don't think this will work.

You mean this one?

Mitesh Rathod 47:36

No, no, no. It is keywords, but it is this is star arts. So no, no, I meant that only square brackets only.

Tirth 47:46

Yeah, and then we can try here. Episodes, total episodes.

I I think we'll just have to play around. I was just curious like if we can have two parameters coming through.

Tarun Jain 48:08

Lord.

Tirth 48:12

Yeah.

Mitesh Rathod 48:17

What is wrong with you, Biper?

Code also release the the IDE plugin not code.

Tarun Jain 48:46

You want augment release yesterday only CLI.

Mitesh Rathod 48:48

Yeah.

Tirth 48:49

Even I think codecs also released CLI.

TJ **Tarun Jain** 48:50

See you later.

Mitesh Rathod 48:53

Huh.

Tirth 48:53

Codex also released CLI.

Mitesh Rathod 48:56

Uh uh now like it has CLI, but uh they uh they'll do authenticate with the JGB plus now.

Tirth 49:05

Oh, they have this row if you see the option 2.

But it is with apply.

TJ **Tarun Jain** 49:10

Uh, so apply. I'll come to apply. I'll tell apply there there is a particular item.

Tirth 49:15

Map would also work. No map would also work, but it is just saying that it is a list and then you don't have to use star star args, just the arg and what you did should just work fine with at least one option 2.

TJ **Tarun Jain** 49:24

Where?

Huh.

Tirth 49:30

Yeah, so row row total episodes greater than 100.

TJ **Tarun Jain** 49:35

No apply. I'll tell you how to use apply because there is a what you call. There is a code that I've written for apply as well.

TJ **Tarun Jain** 49:44

So typically what happens is let's suppose you have that Twitter data that we spoke about.

Tirth 49:48

Mm-hmm.

TJ **Tarun Jain** 49:50

Right in Twitter data I have a column called text.

Tirth 49:53

Mhm.

TJ **Tarun Jain** 49:57

So what are the normalization technique you can apply? Do you remember this keyword?

Tirth 50:06

But.

TJ **Tarun Jain** 50:07

What are the different normalization techniques?
This is something in NLP that we did. It will be a revision.

Tirth 50:18

OK.

Mitesh Rathod 50:20

No.

Tirth 50:23

OK.

OK.

Tarun Jain 50:23

Converting into lowercase.

Tirth 50:28

OK. And then we to make the general synonyms like it converts running to run, what do we call it?

Mitesh Rathod 50:31

Oh.

Uh, removing the stop.

Yeah. No, sorry. Thank you.

Tarun Jain 50:42

Bringing into rote words.

Tirth 50:43

Yeah, bringing in the root words.

Then there was this, you know it would cut it down like fly to FLI or flying to FLI, but that is limit I think OK.

Tarun Jain 50:54

That is fine. Both are stemming and lemmatization, but stemming usually will not use because that doesn't add any grammar meaning. So lemmatization is 1, lowercase is 1 then.

Mitesh Rathod 50:57

That would exist.

Tirth 51:01

Bye.

Mitesh Rathod 51:01

OK.

Tirth 51:04

Right. Limit.

Mitesh Rathod 51:05

Mhm.

Tirth 51:07

And and we remove the stop words.

Tarun Jain 51:08

In NLTQ we saw that.

Uh, stop words.

And then pre-processing. So let's suppose if there is any URLs or anything we have to remote. So now if I display my DF, let's suppose.

Tirth 51:22

Bye.

Tarun Jain 51:27

Anime names is there. You have capital, you have capital. I want to convert this into smaller case. So what I will do is I will define to lowercase.

Tirth 51:33

Um.

Tarun Jain 51:39

And then some entry I'll tell text.

Tirth 51:43

Mhm.

Tarun Jain 51:44

Whatever text I'm getting here.

Mitesh Rathod 51:47

No.

Tarun Jain 51:48

I will convert that into lowercase.

Now DF of.

Mitesh Rathod 51:56

And many means.

Tarun Jain 51:58

Any minute.

Equals to. I'll just copy this, then dot apply. I just have to add this function at the end of the day. What is Lambda?

Mitesh Rathod 52:07

OK.

OK.

Tarun Jain 52:14

What is this trying to do? This is a right? So what is apply taking? Apply is also taking a function which I did here. So apply takes a function. Whatever you are defining here, right? Either you can have it in a different function or you can define it in a single line.

Tirth 52:16

It is an inline function.

Mitesh Rathod 52:20
Um.

Tirth 52:20
Yeah.

Mitesh Rathod 52:24
OK.

Tirth 52:31
Hmm.

Mitesh Rathod 52:32
Mm.

TJ **Tarun Jain** 52:33
So it's the same thing. So play works.

Tirth 52:36
Why not map? It would be one and the same. The map will also do the same thing.

TJ **Tarun Jain** 52:40
Huh. I have what you can use.

Tirth 52:42
Yeah.

Then what is the difference like just thinking out loud?

TJ **Tarun Jain** 52:49
OK, I don't see much difference in apply and map because scope of apply a map. What they will do is they will take a function as a parameter and whatever logic you define in that function it will just apply that to the.

Mitesh Rathod 52:53

Different approach.

Tarun Jain 53:07

Uh, specific functions. So map is the right map and Lambda are Python functions.

Tirth 53:10

Hmm.

Mitesh Rathod 53:16

Mhm.

Tarun Jain 53:18

So if you write map.

If I do map, there is a function called map in Python.

Tirth 53:26

Mhm.

Tarun Jain 53:28

Same goes for Lambda. These are Python functions whereas apply.

Tirth 53:31

OK.

Tarun Jain 53:34

This is a Panda's method.

Tirth 53:36

OK.

Tarun Jain 53:39

OK, so whatever applies there, it belongs to Panda's data frame. So they just added

additional thing for map, whereas map and Lambda are pure Python functions.
So coming back to the previous question, did it give the response?

Mitesh Rathod 54:01
Oh, you.

Tarun Jain 54:04
No, it complicated.

Mitesh Rathod 54:07
Um.
So like apply will work in this case. So in tangibility it says that apply will also in our data frames and non iterator. I mean the objects from data pandas.

Tarun Jain 54:22
Supply will work.

Mitesh Rathod 54:32
Whereas the map will only work on intervals.

Tarun Jain 54:36
OK, wait, it did something. DF you're taking two columns, you apply and what? What is this tuple?

Mitesh Rathod 54:40
What?
M.

Tarun Jain 54:45
OK, tuple is nothing but a data type.

Mitesh Rathod 54:46
Uh.
If you can just check the like a function doc on the apply.

Tarun Jain 54:50

Then.

Oh, OK, OK, OK, OK, I got it. So this is what you said, right?

Tirth 54:57

Yeah.

Tarun Jain 54:59

So let's try this.

Mitesh Rathod 55:09

Uh.

Tirth 55:11

Int object is not subscribable.

Tarun Jain 55:16

So here it's applying tuple.

And then tuple.

Mitesh Rathod 55:23

So it it works for apply because it it understands data free, but it won't work for map because map only understands it drivers.

If you can, uh, convert it into the installer.

Tarun Jain 55:42

Uh, for which line? This one?

Mitesh Rathod 55:44

Yeah.

Tirth 55:45

Yeah, so.

TJ **Tarun Jain** 55:47

So what did you say, Lambda?

Tirth 55:49

Let it be X and then typecast X to list in after X, yeah, and then we can try.

TJ **Tarun Jain** 55:55

OK.

Tirth 56:01

Maybe.

TJ **Tarun Jain** 56:04

So here probably.

Tirth 56:04

In object is not iterable it it should be total episodes.

Mitesh Rathod 56:08

No. So like maybe uh, I think.

TJ **Tarun Jain** 56:15

No. So basically what is happening is it will take the entire total episode as zeroth index. So if I do 0.

Mitesh Rathod 56:16

Music.

Index.

In the object is not.

So what is the value we are getting in the X? Can we do that?

TJ **Tarun Jain** 56:37

That's a tricky thing.

So if I apply.

Tuppal of Axis 1.

So bad.

Then where if I print?

Mitesh Rathod 57:00

Mm.

Tarun Jain 57:01

Oh, then map.

Lambda of 0th entry X of 0 if it is more than 100.

Tirth 57:15

Hmm.

Mitesh Rathod 57:15

Then wrap it to the to DS.

Tarun Jain 57:19

But why you go with this approach? This is one line and simple.

Tirth 57:24

Yeah, yeah, no, no. I was just thinking if we have multiple variables.

Tarun Jain 57:27

Yeah, what you put?

Tirth 57:30

Yeah, that's fine. Like I it was.

Tarun Jain 57:30

There is an up, but it's not.

Tirth 57:34

I I tried GPT and didn't got an answer. I just dropped it here. So you know if you knew something from top of your head, but something important, you know we can continue.

Mitesh Rathod 57:37

It.

Tarun Jain 57:42

Cool.

Tirth 57:43

Yeah.

Tarun Jain 57:48

Where where we apply also we already covered now, so you understood right how apply works. So if you want to modify any columns for normalization in order to train the model, define a function for it. Here I can also have for remove stop words.

Tirth 57:52

Yeah.

Yes.

Tarun Jain 58:07

And whatever we covered in NLTK, right? So you can apply that logic and then return true or false.

Tirth 58:16

Hmm.

Tarun Jain 58:16

So if that particular, what do you call? If that particular word exists, then fine. If not, just remote. If not, we can do one more way. Pass text then.

Tirth 58:26

Return text dot replace.

Tarun Jain 58:28

Whichever text word is there, right? If it is within the stop words, we will not append it to our final string and then you can return string itself or cleaned text.

Tirth 58:30

Yes.

Mhm.

Tarun Jain 58:45

Logic of stopword.

Same goes for lemmatization. Same goes for removing hyperlinks. Once you define this logic, tell which column you want to append or modify, add, apply and then to lowercase.

So apply is done, map and Lambda is done. Single filtering and multi filtering is also done. Any questions here?

Tirth 59:12

OK.

Tarun Jain 59:14

So once again I'll just show the logic of multi filtering because this is very important.

DF this is same. You first start with DF then start with your first condition which is.

Mitesh Rathod 59:30

Interple.

Tarun Jain 59:30

I need a data frame whose total episode is more than 100 and then it's a tuple and then this particular entire condition I want to apply bitwise and and check with another condition and then it will display.

Whatever condition is there, if it is true, it will display it.

And with map function it's also same. When you're using map function, you don't have to define it inside a.

Parenthesis.

Is this clear?

Mitesh Rathod 1:00:03

Yeah.

Yes.

Tarun Jain 1:00:05

OK, so there are a few more comments. What if you want to print the maximum in your data set? It's the same logic. I'm starting with DF and then if DF of total episode matches with the DF of episode which is Max.

Let's suppose I print this separately.

It will be 1014. Now what I'm trying to do is I'm checking all the entries in my data. If it is equal to Max, I want to print that entire row.

Tirth 1:00:32

2.

Mhm.

Basically it works on series of true and false. So whatever condition we write in this, it is a series of true and false. Everything else will be false and only one row will be true, so that will be printed.

Tarun Jain 1:00:50

No, correct.

So if you remember in series, how did we work in series? First you wrote a logic, you upended true or false. Once you add that particular series, at the end of the day, this is series which is only one column. Once you have that column, you're only adding that in a DF.

Tirth 1:00:53

Yeah.

Mhm.

Mm.

Tarun Jain 1:01:09

Once you add that in DF, it will filter it out. If it is false, it will remove. If it is true, it will keep it.

Tirth 1:01:10

Mhm.

Mm.

Tarun Jain 1:01:16

Same goes with minimum. I'm adding DF total episode then dot min whichever is minimum you'll have that entry. So in sales also what you can do is you can have specific column and check which is the minimum invoice that was generated or which was the maximum invoice that was generated.

Tirth 1:01:34

Mm.

Tarun Jain 1:01:41

OK, these are still the same commands. There is no new here min Max min. OK, so now this concept is very important. The correlation. I've written too much details over here, but I'll summarize what will happen. Let's suppose you have 5 columns.

Two columns.

Are have very close value.

This is mainly, uh, let's suppose the data is related to our disease.

And can anyone tell me what is usually added in RDCS data?

Mitesh Rathod 1:02:29

I did it.

Tarun Jain 1:02:30

Hey, I mean, uh.

Oh, you have it. You have under.

Tirth 1:02:38

Why? Why would you have heart disease? So age, gender?

Tarun Jain 1:02:40

Then you have. Then you'll have.

Mitesh Rathod 1:02:43

Sugar, sugar, HBL, LDL, blood pressure, blood glucose.

Tirth 1:02:48

Let's not smoke, you know.

Tarun Jain 1:02:51

Good.

BP.

Tirth 1:02:54

It is a change number.

Tarun Jain 1:02:59

OK, so we added more than that itself, but that's fine. So now let's suppose 2 columns are there, which is cholesterol and.

Which one do you think goes hand to hand with cholesterol? Is it sugar?

Tirth 1:03:12

BP.

Mitesh Rathod 1:03:13

Blood pressure.

Tarun Jain 1:03:13

So now when you're building a predictive maintenance algorithm model, let's suppose you're building a model and you know right all this columns is nothing but your features.

Tirth 1:03:16

Yeah.

Hmm.

Tarun Jain 1:03:30

Now when you add any feature to your model, everything is dimensions.

And our goal is usually.

To reduce.

Curse of.

Dimensionality.

So what we have is we have a concept called correlation. In correlation, if there are any two columns which are very close in the sense whatever values you have, it is correlated to each other. For example, if it is very close to one, that means they are positively correlated.

If they're positively correlated, what you can do is you can drop one column and you can only keep one column.

Mitesh Rathod 1:04:15

Yeah.

Tirth 1:04:16

Hmm.

Tarun Jain 1:04:16

For example, college store and BP. Both of them have very close values and it's not making sense to keep both of them because they are highly correlated. So when any two columns are highly correlated, when you're building any model, it's fine to pick only one and drop the another one.

And how do we identify it if the columns 2 columns are positively correlated, which is very close to 1.

Tirth 1:04:41

We divide 1 column by another one and we see if it is near to 1.

Tarun Jain 1:04:42

Yeah.

Huh. You'll usually get a matrix which will tell OK this column and this column what is the value of it? Where did we see this earlier for cosine similarity we saw if you remember our diagonal was one.

Tirth 1:04:51
Yeah.

Mitesh Rathod 1:04:59
Yes.

Tirth 1:05:02
Hmm.

Tarun Jain 1:05:02
So here also what it will do is it will create a metrics like that for two entries. What is the correlation between these two columns?

Mitesh Rathod 1:05:03
This.

Tarun Jain 1:05:12
And if it is close to one, it's fine to remove one of the column and keep only one and then we have two columns.

Tirth 1:05:19
M.

Tarun Jain 1:05:25
Like age and sugar.
And these are negatively correlated, which is -1.
Let's suppose I have -0.6. These are negatively correlated. That means if sugar is increasing, the age is very less or the age is more, right? There is no correlation between these two. During that time we can keep both the entries.

Tirth 1:05:49

Mm.

Tarun Jain 1:05:50

And the third thing is the uh null.

Correlation if it is very close to 0.

You can keep that as well. You can keep both the columns.

Mitesh Rathod 1:06:02

No.

Tirth 1:06:04

OK.

Tarun Jain 1:06:06

So the range is -1 to 1 where one is positive correlated.

Mitesh Rathod 1:06:08

OK.

Tarun Jain 1:06:16

-1 is negative correlation.

Mitesh Rathod 1:06:18

Like.

Tarun Jain 1:06:21

So when I usually give talks right related to correlation, I used to give an example. Let's suppose you're sitting in exam hall. If two people are copying and writing any example or writing any exam, they're positively correlated. So what the frontbencher will write? The backbencher will also write the same answer.

And negative correlation is like you're not copying, so doesn't matter what the frontbencher score is, the backbencher score might be different. So that's positively correlation and negative correlation. Is this clear?

Mitesh Rathod 1:06:52

Yeah.

Tirth 1:06:54

Yeah.

Tarun Jain 1:06:56

OK, so since our data, whatever we had so far, there is no much of the correlation or not because if you see there is too much of text in which data set was this movie data set and in animate data set also I just had total episode which is just one column so it is not.

Tirth 1:07:05

Mm.

Tarun Jain 1:07:15

Easy for me to identify the correlation. This is where you have one more data set which is random dot CSV. So now what we can do is we can read that particular file because that particular column has more numeric data.

Tirth 1:07:20

Hmm.

Tarun Jain 1:07:29

So if you print numeric dot info.
So if you see you have float, float, float, there is too much of data here.
And in actual data set you will have more than 40 columns. You will never have less than 30 columns. If you have any data which is less than 30 columns, that is just experimental data. In actual scenarios you will usually have more than 3040 data.
I mean columns.

Tirth 1:07:59

M.

Tarun Jain 1:08:01

Now if you look at for describe, there are too many entries where you have described. Earlier we just had for one or two entries only.

So now if you do correlation.

Did I get an error?

Numeric.

Mitesh Rathod 1:08:28

So.

Tirth 1:08:29

Could not convert string to float M.

Tarun Jain 1:08:44

Diagnosis I'll remove.

Numeric dot drop.

How do we drop? You remember this syntax, right? So if you want to remove any column, just define that column name axis one, and if you want to permanently remove it, just in place equals to true.

Mitesh Rathod 1:08:58

Yes.

Tarun Jain 1:09:09

So now if I check info.

I have only float.

Huh. So if you see the diagonal, it is 11111. That means ID and ID will always be correlated, right? Because the values are same.

Tirth 1:09:30

Thanks.

Tarun Jain 1:09:32

And which is highly correlated with ID. Obviously nothing should be there because ID is a sequential number. If we check for the entry, if you see area mean and perimeter mean it is.

Tirth 1:09:37
Hmm.

Tarun Jain 1:09:46
0.98 it is highly correlated. So what will I do? I will only keep one of them instead of both.
You understood?

Tirth 1:09:55
Yeah.

Tarun Jain 1:09:56
Here also if you see concave points mean and perimeter mean, they're highly correlated. So this usually takes too much of time. It's very easy concept, but when you preprocess right?

Tirth 1:10:05
Um.
Uh.

Tarun Jain 1:10:12
Building model is just one day task, but what column to choose? What column to reject? It usually takes two to three months just to figure that out.

Tirth 1:10:14
Mm.
Mm.

Tarun Jain 1:10:24
So most of the anomaly detection use cases that we take, right?
Right.

Anomaly detection where you have to tell whether this particular entry is flagged normal or not flag normal.

Mitesh Rathod 1:10:31
OK.

Tarun Jain 1:10:41
Or trigger any alert.
We can't randomly remove any columns. So if we remove any column, we need to have a proper justification why that was removed and then we create a documentation of what new columns were created, what columns were removed, and then we give a justification of why that column was removed.
So to prepare that document usually takes three months of time and it seems easy, but if you see.

Tirth 1:11:01
Mhm.

Tarun Jain 1:11:10
Now if you look at perimeter mean, area mean and concave points mean out of these three, which one to pick?

Mitesh Rathod 1:11:19
Mhm.

Tarun Jain 1:11:20
Right. It's easy to say area mean, but what if I wanted concave mean? Even though perimeter mean and area mean are very correlated to concave, I can keep concave and then I can remove area mean. So those reasons it's very difficult to justify.

Mitesh Rathod 1:11:25
Oh.

Tarun Jain 1:11:37

And this is just one technique. We have other techniques as well where you have something called as chi square.

Tirth 1:11:46
Mhm.

Tarun Jain 1:11:47
Or F1.

Mitesh Rathod 1:11:49
Yes.

Tarun Jain 1:11:51
So we perform all these things and then we come up with the report. So correlation is the simple 1C square. Again, it's very complicated. We have to use SK learn. I hope everyone knows what SK learn is.
We used it for uh from scikit learn.

Mitesh Rathod 1:12:08
Yeah.

Tirth 1:12:08
Yes, yes.
Yes.

Tarun Jain 1:12:14
So it was this is scikit-learn.
And then when you import it is SK learn dot feature engineering we imported TFIDF.

Tirth 1:12:29
Yeah.

Tarun Jain 1:12:30
And then from pair wise, sorry it was metrics dot pair wise.
We import cosine.

Similarity. So yeah, this is something that most of the data scientists spends too much of time. It's like 3 months, two months just to identify what what columns to keep, what columns to remove. And this can go too much in depth, but the basic one is the correlation.

Is this clear?

Mitesh Rathod 1:13:01

You guys.

Yeah.

Tarun Jain 1:13:05

OK, we'll quickly move on with the matplotlib. This is it from pandas. The most important is the filtering in pandas.

Mitesh Rathod 1:13:08

Thank you.

OK.

But.

That'll be.

Tarun Jain 1:13:21

So I'll open this URL again. Are you able to see this GitHub?

Mitesh Rathod 1:13:27

Yeah.

Tarun Jain 1:13:29

So over the weekend what you can do is you can also test out this Part 2 pandas. It's same what you have done in part one. The only thing is you're applying more filtering, but the logic of filtering is same.

Mitesh Rathod 1:13:38

So.

Tarun Jain 1:13:44

And if you see you also have uh fill and name.
So we saw this command earlier as well. Now what is drop and name? You define an axis one and then in place true. So this is a very danger command. So what this will do is if it finds any entry which is empty in that column, it will remove that column itself.

Mitesh Rathod 1:14:09
That's OK.

Tarun Jain 1:14:11
So use this by cautious, right? What we used was we used drop.
Where was that comment?
Correlation head info. We used numeric drop and we know what column to pick and then I'm defining axis one and then in place to be true here when it comes to.
Drop and it's like drop everything which is empty.
And you have different ways to do it. One is you can define certain threshold. If it is more than three then you can remove it or this value can be more as well. You can keep this value more. So usually we prefer this syntax.

Mitesh Rathod 1:14:55
M.
Yes.

Tarun Jain 1:15:03
So you have the code available. Most of the commands that you see in the Part 2, we have already covered that.
So we can click on this matplotlib.

Mitesh Rathod 1:15:22
Mhm.

Tarun Jain 1:15:29
And download this particular file.
And if you also open our own GitHub Red, the Python AI workshop, here you have notebooks. Inside notebook you have EDA and if you see you have pokémon EDA,

then you have video game sales.
So I've already added two examples and the specific CSV file for that. Are you able to see this?

Mitesh Rathod 1:15:52
OK.
OK.

Tirth 1:16:06
Yes.

Mitesh Rathod 1:16:06
Mhm.

Tarun Jain 1:16:07
So this is the same repo that we are maintaining for all the code snippet as well for tokenization, quick start Python in EDA. As of now we have not covered it yet, which we will cover today and OK, we are already close to the time but.
If you look at this thing, numpy, pandas, matplotlib, numpy you have pandas code we just completed. Here also you have the same notebook which is matplotlib notebook file. So once you download come back to Colab.

Mitesh Rathod 1:16:39
Uh, which way you can download?
From like 100 days or uh from your yeah.

Tarun Jain 1:16:46
Both are same.

Mitesh Rathod 1:16:52
Mhm.
Yeah, 100 is.

Tarun Jain 1:16:59

So now come back to Collab, click on upload, browse.

And then matplotlib.

Mitesh Rathod 1:17:13

Good.

And.

Tarun Jain 1:17:21

So why was Nampa used?

Mitesh Rathod 1:17:28

Calculations.

Tarun Jain 1:17:31

For vector calculation.

Or data creation. So this data is nothing but your Arish.

Or you can also say vectors. What about pandas?

Mitesh Rathod 1:17:48

Manipulation.

Tarun Jain 1:17:49

Data manipulation and now we have matplotlib.

Which is mainly used for data visualization.

And these two, if you combine, you can build sales agent kind of use cases, right? Because in sales agent pandas and matplotlib is your tools, right? You upload your data, you tell agent to write pandas and matplotlib code once it's write that code. You will execute that in interpreter, get the result and then give it to the agent. So what did these two happen? These two are your tools, right? So when you build sales agent or anything related to tabular column.

Mitesh Rathod 1:18:30

Yeah.

Tarun Jain 1:18:38

Which is your data analysis agent. These two libraries are very important to understand, pandas and adlotlib, even though this is data science framework related, but in order to build better agent use cases.
It's better to learn this too.

Mitesh Rathod 1:18:54
OK.

Tarun Jain 1:18:56
And what is this syntax to install it? You can directly do pip install matplotlib, but since we're using Colab, we don't have to install it.

Mitesh Rathod 1:19:03
Yeah.

Tarun Jain 1:19:07
So the fundamentals are every single plot, every single.
And when I say plot, you can treat this as chart or you can also treat this as diagram.
OK, it has something called as figure and axises.
So what is this figure? It's nothing but the empty white screen. And what is the accesses? It's this 10 to 10 to 1.

Mitesh Rathod 1:19:34
The.

Tarun Jain 1:19:43
And you can also increase the figure size. Let's suppose you have this default. OK, let's run it.
So this syntax is import matplotlib. Then you have a function called pyplot. This is nothing but Python plot. This is how you import matplotlib.
Now since this is very big right, every single time I need to use this. So what we usually do, we define yes and PLT. Now every time I define PLT it will use matplot matplotlib dot pyplot. This syntax is clear right?
What this ES is doing?

Mitesh Rathod 1:20:23

Yes.

Tarun Jain 1:20:23

This is the short form PLT short form of matplotlib.pyplot.

And now the basic syntax of creating any diagram is you have a figure and you have a axises. So the axis is nothing but your Y axis and your X axis. This is X, the bottom, the horizontal one. The vertical is your Y axis.

And PLT show it is to display. This is to display.

So once you define these two, you just want to display it if I remove this.

It's displaying in collab, but if you run this in VS code it will not display.

You can try these two lines in Colab, I mean VS code.

Tirth 1:21:13

But is it? Is it a canvas display? Like where would it display?

Tarun Jain 1:21:18

So when Collab is interpreter, once you define this PLT figure, it's adding this particular diagram.

Yeah.

Mitesh Rathod 1:21:26

It's an image, right?

Right.

Right.

Tirth 1:21:30

OK.

Tarun Jain 1:21:32

So this is the actual output.

So this is the size of your image.

Mitesh Rathod 1:21:43

M.

Tarun Jain 1:21:47

640 is nothing but your width. 480 is the height.

Mitesh Rathod 1:21:51

OK.

Tarun Jain 1:21:54

Now we can also increase it. Now what I'm trying to do is the syntax is same. When you do PLT figure, I'm defining the figure size where I'm making it 15 and 12:15 is in the sense you're increasing your width. 12 is nothing but you're increasing the height. So the first parameter is always width.

Comma type.

And then this is same PLT dot access and then PLT show. Now you have a very big diagram.

Now where is this used? If you want to display, let's suppose your data as images.

And you want to display bulk images.

That is, you want to display 4 images, not four. Let's suppose I want to display 12 images in four cross 3.

That means I have total 4 rows, 3 columns with total 12 images. During that time you will need a very big figure size to display that particular image.

Tirth 1:23:02

Mm.

Tarun Jain 1:23:04

So the syntax is familiar. You just have figure and accesses. Now we'll go with different kind of charts.

the

And you can also add a grid if you need. Grid is nothing but it will just add a box for every single entry and that is nothing but AX dot grid on.

Mitesh Rathod 1:23:35

No.

Tirth 1:23:52

OK.

Tarun Jain 1:24:08

Are everyone able? Uh, I mean, are you guys able to run this?

Mitesh Rathod 1:24:11

Yes, yes.

Tirth 1:24:11

Yes.

Tarun Jain 1:24:14

OK, figure access and if you want grid you can add it, but usually no one will add grid.

OK, so now we will start off with line 9 line plot. So most of this data set that you have right when it comes to stock prices, what kind of chart is that?

Mitesh Rathod 1:24:29

M.

It's a line job.

Tirth 1:24:42

Can you delete stock?

Tarun Jain 1:24:42

Uh, Binance.

Mitesh Rathod 1:24:44

Uh, kind of like.

Tirth 1:24:45

Candlelight, Yeah, Candle, Candle chart.

Tarun Jain 1:24:53

OK, you are saying uh, the particular uh, this is possible.

Tirth 1:24:55

Yes. And they'll stick chart, yeah.

Tarun Jain 1:24:59

So if you see here whatever you have this right this specific thing, this is called box plot.

Mitesh Rathod 1:24:59

Yeah.

Tirth 1:25:05

OK.

Mitesh Rathod 1:25:11

Post.

Tarun Jain 1:25:12

So we'll start with line line plot and then we'll proceed with few more plots and one of the common thing we usually use is we use something called as heat maps as well. Heat maps is like shows some of the pattern.

If it doesn't make sense, for example, let's suppose I have any empty data and I want to display a heat map of it. So wherever your data is completely filled, it will not show any heat map. Wherever there is empty, it will add certain color. Hey, in this particular entry there is an empty data.

So I hope everyone remember this syntax.

Tirth 1:25:52

In space.

Tarun Jain 1:25:54

What does line space do?

Tirth 1:25:56

So 0 to.

Tarun Jain 1:25:59

It will go zero to 15.

Tirth 1:26:00

In 150 steps.

Tarun Jain 1:26:02

High in 150 steps.

So if this is 0, then there are some decimal points.

Tirth 1:26:08

Yeah.

Mm.

Tarun Jain 1:26:13

And then what I'm trying to do is I'm displaying a sine graph which is sine wave and if I want to use sine I will directly use it from.

Mitesh Rathod 1:26:16

OK.

Tirth 1:26:23

OK.

Tarun Jain 1:26:24

Num PY and for line plot for line chart we just have to use PLT dot plot. So PLT dot plot is for line.

And here what I'm trying to do is this is the syntax. If I want to add label, I can define label equals to sine wave. Color is green. Same goes for NP cos XI want to define a label. Label is nothing but cosine wave.

And then color is orange. You can keep it red if you just add R it is red and then PLT legend is nothing but whatever box you see right that is legend.

Mitesh Rathod 1:26:56

Nothing anything.

So.

Tirth 1:27:06

Yeah.

Tarun Jain 1:27:07

And you can also add it as lower left.

Did you want to add this particular uh legend?

Mitesh Rathod 1:27:13

But.

Tarun Jain 1:27:16

Thanks.

So plot is nothing but line, then display the sign graphs label. Whenever you add label that's specifically mean that you are displaying 2 different graph in a single plot. Right during that time you will label it. If not, usually label is not defined. If you define label and if you want to let user know what different labels have defined, we use legend. So legend you can tell which location you need, whether you need upper left.

Mitesh Rathod 1:27:37

Good. Yeah.

Tarun Jain 1:27:51

Lower left, upper right, lower right.

And you can also increase the size of it. How do we increase it PLT dot figure?

Mitesh Rathod 1:28:05

Yeah.

Tarun Jain 1:28:09

Pick size equals to 10 comma 10.

Mitesh Rathod 1:28:12

Mhm.

Tarun Jain 1:28:20

Is this clear?

Tirth 1:28:24

Yeah.

Tarun Jain 1:28:28

We'll also try this on the actual data set when we perform EDM.

And this is the common syntax. If you display KK is nothing but black, R is nothing but red, B is nothing but blue, G is green, Y is nothing but yellow. These are the short form command whenever you're defining color. For example if I do K here.

Instead of red, it will become black.

And you also have certain markers, so let me define a marker.

Mitesh Rathod 1:29:03

Sit.

Tarun Jain 1:29:07

And here what I'll do is I'll just add plus.

If you see here it is adding plus plus. If you need dotted you can add dotted line.

Mitesh Rathod 1:29:20

Sing.

Tarun Jain 1:29:23

What is the syntax for dotted?

Mitesh Rathod 1:29:27

It's underscored or not?

Tarun Jain 1:29:33

No under score won't work. OK, it's under score, but I wanted a dotted line.

Mitesh Rathod 1:29:39

Super dot test.

I have a full stop.

Tarun Jain 1:29:48

It's starting on that particular line only.

And you also have plus, minus, then you have star.

Mitesh Rathod 1:29:56

M.

Mhm.

Tarun Jain 1:30:01

So what this will do is it will just denote hey this is the entry. This entry is plotted here, then the second entry is plotted here, then third entry is plotted here.

And that is defined by marker and the common syntax for marker is you have plus dot then this is not under score, this is hyphen, then you have star, then you have bracket and if you just define it as D it will add diamond.

Is this here?

Tirth 1:30:39

Mm.

Tarun Jain 1:30:41

Again, I'm adding one more example. I'm going from zero to 15. Then you have sine wave, then color is red. Here color is blue and when I say marker is edge, that means this is hexagon. D is nothing but diamond.

Mitesh Rathod 1:30:44
It.

Tarun Jain 1:30:58
And legend it's upper right and then PLT show PLT show if you are using VS code. If you're using collab, it will directly display the image.
Now what if you want to define a title for this? Y axis you need a title, X axis you need a title. Again we are taking the same example. These two are same.

Mitesh Rathod 1:31:13
4.
It.

Tirth 1:31:19
And label.
Hmm.

Tarun Jain 1:31:23
Legend is also same. Then you're displaying a title. The title is nothing but trigonometry function X label. You're defining it X. Then Y label is nothing but sin X and Cos X.
So this is mainly used when you're writing any papers or if you want to display any charts. During that time you will use this 3.

Tirth 1:31:43
Mhm.

Tarun Jain 1:31:47
Functions PLT title, PLTX label, PLTY table, matplotlib. If you know what functions to use, it is just PLT dot. That's it. There is nothing related to you define a variable. So usually what happens in pandas.
You define a variable, then you write something, then you have attributes right? Like

it can be ndim, then var dot shape. It can be anything. Same goes for pandas as well.
But when it comes to matplotlib it is just PLT and what you need, that's it.
Let's take an example of bar plot as well. I'll remove it. I don't know why you have imported it multiple times.
Now you have two entries. You have one piece, you have Naruto, Bleach, Gintama, Fairytale, and then you have one more column right now. What you need to do is you need to display a bar plot whenever you're displaying a bar plot, whatever you're displaying in the first.
Variable. This is your X axis.
And whatever you're displaying on your second variable, that's your Y axis. Can you see it?

Mitesh Rathod 1:32:59
Yeah.

Tirth 1:33:00
Yes.

Tarun Jain 1:33:00
It's a list. Whatever you have in the second, it's your episode, whatever I have at first. It's an aluminium.

Mitesh Rathod 1:33:10
That is so.

Tarun Jain 1:33:12
And then you can define your label. I'll add title as well.
Do you think Legend will work here? Do we need Legend?

Mitesh Rathod 1:33:33
We can do.

Tarun Jain 1:33:35
So if I add up, will this work?

Tirth 1:33:39

It should.

Mitesh Rathod 1:33:40

Yeah.

Tarun Jain 1:33:40

It should not because there is.

Tirth 1:33:43

OK.

Tarun Jain 1:33:45

No, it has randomly added anything. There is no what you call label.

Mitesh Rathod 1:33:49

Yeah.

Tirth 1:33:49

That's it.

Tarun Jain 1:33:50

If you see here here, why did it add sine wave and cosine wave? Because it add label. If you don't define a label.

Tirth 1:33:54

Hmm.

Mitesh Rathod 1:33:56

OK.

Tirth 1:33:59

Hmm.

Tarun Jain 1:34:00

It will randomly put something.

So now you will never know what is this purple color and what is this blue color. So whenever it comes to line plot, having label is very important. But when it comes to bar plot, you already know OK, this is your X label, it's already labeled and then.

Mitesh Rathod 1:34:15

OK.

Tarun Jain 1:34:21

This is your Y axis.

So if you see here, it's already labeled in bar plot, but in line plot it's never labeled.

Tirth 1:34:29

Mm.

Mitesh Rathod 1:34:30

Absolutely.

Tarun Jain 1:34:31

But why it added you? Let's suppose if I use lower, will it add L?

Tirth 1:34:36

Yes.

Tarun Jain 1:34:37

Uh, it's taking the first variable.

Tirth 1:34:39

He.

Tarun Jain 1:34:41

But it's not adding at lower right if you see.

Tirth 1:34:43

Yeah, yeah.

Mitesh Rathod 1:34:44

Yeah.

Tarun Jain 1:34:46

Lower left.

So legend will only work when you have labels and bar plot has default labels.

Mitesh Rathod 1:34:54

Maybe.

Tirth 1:35:01

But.

Tarun Jain 1:35:02

And you can also change the rotation of it. Let's suppose as soon as the entry increases you will have multiple labels over here. During this time what you can do is PLT dot you have something called as X ticks.

And here we usually define rotation.

Hi if you see there is a variable called rotation.

And here I'll just define 75 degree. So now you have 75 degree. So now whenever you have multiple entries, if there is any label which is big, it will not consume more space.

So if you don't add this rate.

There might be chances once more labels is increased, whatever the size of blue is there, that will become very less.

Mitesh Rathod 1:35:54

OK.

Mm.

Tarun Jain 1:36:01

And during that time, whatever name is there, if it is big, if it has at least three variables, it will consume too much space like JoJo's Bizarre Adventure. So just imagine if you add JoJo's Bizarre Adventure, how big name that is. So you usually put a rotation over here.

Tirth 1:36:15

Mm.

Tarun Jain 1:36:21

Is this clear? Bar plot is very simple. Uh, whatever you need to.

Tirth 1:36:22

Yep.

Tarun Jain 1:36:27

Map it against your first variable will be X, your second variable will be Y.

And now we have histograms. Histograms is basically on distribution level. Let's suppose 130 is there. How many times was 130 occurred? It was occurred one time, right? So when do you usually use it if you want to count any values and the probability, not probability.

The overall count during that time you will usually use histogram. For example, if you see 190, how many times?

Mitesh Rathod 1:36:56

This.

Tarun Jain 1:37:06

Not 190. If you see 185 to 190, how many values are between this range? That is the count which is 3. So in my given data I have total 3 entries which are between 185 to 190.

Is this plot here how histogram works?

Tirth 1:37:27

Mhm.

TJ **Tarun Jain** 1:37:28

So you have ites and you're randomly giving some ite names and then you're defining PLT dot hist. Hist is nothing but histogram and then you have to pass your weights.

Mitesh Rathod 1:37:29

You.

TJ **Tarun Jain** 1:37:49

If I make this bins as 40, the size will increase.

If you see here now the size has decreased and there is too much of entries. This is very specific. Now for 130 it shows one. Now what will be this 11414142143? Roughly it will be around 146 or 147.

Mitesh Rathod 1:38:09

Mm.

I want see.

TJ **Tarun Jain** 1:38:15

It's 146.

Histogram is for the count.

Bar plot is when you have two different entries and you're ranking them against them that during the time you'll have bar plot. Line plot is nothing but you have sequence of digit and you just want to display it.

So line plot, where will this be used? I'm coming with an example.

So in cricket matches, what kind of plot do they use?

In in cricket matches it is bar plot.

So what is your X axis? X axis is nothing but overs. Y axis is nothing but the runs.

So this bar plot we can give multiple examples. Histogram. Let's suppose you have.

Mitesh Rathod 1:39:13

Weather.

Two days we get some certain.

Tarun Jain 1:39:17

Entries of height and you want to count right? Whenever the option is count, the best option is histogram line graph. I'm not able to come up with an example.

Uh, the benchmarking scores and sales. Let's suppose in January what was the sale?

You have bar plot but you are also adding line graph about that. But this is an example.

Mitesh Rathod 1:39:44

Overleaping.

Intersect.

Tarun Jain 1:39:54

Maybe it can be in terms of time period.

Mitesh Rathod 1:40:05

Wait of a person our time.

Tarun Jain 1:40:06

So it's this examples most of the thing wherever you're using line graph, it's over some period of time.

Mitesh Rathod 1:40:14

Sales tablets.

Tarun Jain 1:40:14

So 220 what was the population then 221 and on the left hand side you have certain count and then you're just this is marker. If you see whatever you have here right this is dot this is marker.

And then over the time period of time, you're just telling how the line graph has increased and this can be mainly used for benchmarking as well.

OK.

So let me tell you how in benchmarking, let's suppose you have GEMA model.

Then you have Mr. model.

And then what else do we have?

We have deepseek.

And then when?

So when you're doing benchmarking right, let's suppose I do PLT dot plot.

I will have Gemma's core.

And label will be a last swag. Do you remember this word?

Mitesh Rathod 1:41:28

Yes.

Tarun Jain 1:41:28

Color I will display it as blue.

Tirth 1:41:28

Hmm.

Tarun Jain 1:41:36

Marker.

I'll keep it dot because here also if you see it's dot.

Because here also if you see it's dot.

Now similarly I'll repeat it for.

Other four, instead of Gemma, I laugh Deepseek.

Mitesh Rathod 1:41:51

Sorry.

Yes.

Tarun Jain 1:41:55

Here I'll have Mr. Here I'll have one.

And this will be Deepsik. This will be Mr.

Mitesh Rathod 1:42:09

OK.

Tarun Jain 1:42:11

This will be when and then what will I have? I'll have PLT dot legend.

Mitesh Rathod 1:42:17
OK.

Tarun Jain 1:42:19
Location is nothing, but I'll keep it upper right.
And then PLT show.
But here I'm just adding one single LS flag. Usually what you can do is you can add GEMA score across multiple benchmark and then label it just as GEMA and then you can have plots like this mainly for benchmarking and then.

Mitesh Rathod 1:42:51
Yeah.

Tarun Jain 1:42:57
If there is anything over some period of time, right? If you see product rents by month and if you have multiple products, the desktop, laptops, tablets. So this is something that you can recreate for video game sales also. I hope we should have something.
Where we can use line plot, but I hope these things are clear. This index if you just use PLT plot it is for line.

Tirth 1:43:20
Yeah.

Tarun Jain 1:43:23
PLT bar is for bar plot, then East is for histogram, which is mainly for calculating the count frequency and then you have pie chart. This is very straightforward. I hope pie chart it can be used in most of the cases.
So this is to just tell the percentage distribution of that particular entry. For example computer there is 92. What is the percentage distribution that falls under computer?
And then you have 91908571.
And then if you see this exploded, if I don't give explode, I forgot the syntax of explode.

Shadow is nothing but can you see this Gray shadow Gray color here below every entries?

Mitesh Rathod 1:44:07
No.

Tirth 1:44:15
Yes, yes.

Tarun Jain 1:44:15
Hello.

Mitesh Rathod 1:44:15
Yeah, this.

Tarun Jain 1:44:16
This Gray is nothing but shadow. This auto PCT is nothing but if you see it is going through 3 decimal .729 so.

Tirth 1:44:24
I mean, why is electronics and management out like you know, they're pulled out of the Pi?

Mitesh Rathod 1:44:25
Yeah.

Tarun Jain 1:44:32
It is not because of exploded. If you see 0.2 is nothing but.
Electronics came out, then management 0.15 came out.

Tirth 1:44:40
OK, OK. OK. OK.

Tarun Jain 1:44:43
So only I wanted to remove this and check the result once. Now if I run it.

Tirth 1:44:44

I.

Mitesh Rathod 1:44:47

Yes, yes.

Tirth 1:44:48

Yeah, I can get it, yes.

Mitesh Rathod 1:44:49

Is going through.

Tarun Jain 1:44:51

It is pure circle and you have shadow. If you want to remove shadow you can remove it.

Tirth 1:44:52

Mm.

Perfect. Perfect.

Tarun Jain 1:44:59

And now if I remove auto PCT it will add too much of values.

Tirth 1:45:04

Mm.

Tarun Jain 1:45:05

OK, it's not even adding any value.

Mitesh Rathod 1:45:07

OK.

Tarun Jain 1:45:10

And it's better to add just one.

So what the 100% for that particular individual candidate student is good at computer?

Mitesh Rathod 1:45:20

Yeah, please.

Tirth 1:45:25

Mm.

Tarun Jain 1:45:26

So whenever it comes to percentage distribution, we usually go with pie chart.

So auto PCT it is just to add the label shadow if in case you want to look it better and explode how many entries do I have have total 5 entries. I want to keep computer outside.

Tirth 1:45:33

M.

Mitesh Rathod 1:45:48

I need to.

Tarun Jain 1:45:48

Right, So what I'll do is.

Marks is there.

Marks off.

Mitesh Rathod 1:46:06

Yeah, it's.

Tarun Jain 1:46:07

Max.

Oh.

It's 92. So what I need to do is I need to find the index which is the index of 92 zeroth index. So I'll keep the zero point to outside then 00.

Mitesh Rathod 1:46:11

Is that.

Tarun Jain 1:46:26

And then remaining zero. This I will add in explode.

So computer is outside.

Tirth 1:46:36

M.

Mitesh Rathod 1:46:38

OK.

Tarun Jain 1:46:44

OK.

0.5 looks much better.

Mitesh Rathod 1:46:51

That's great.

Tarun Jain 1:46:53

Is this clear? Line, bar, histogram, pie chart and then you have scatter plot. So scatter plot, whatever you saw earlier, right? Correlation.

Tirth 1:46:55

Yeah.

Tarun Jain 1:47:08

So whenever we use correlation where you are trying to display 2E2 vectors at once, during that time we'll use scatter scatter plot. So scatter is mainly used.

During correlation.

Why? What is happening in correlation? You're comparing it.

So whenever.

We compare.

Two columns.
We can see the distribution.
Using scatterplot.
So for scatter plot also what you can do is you need to have two vectors define X&Y and then you can have any color. If you don't give any color it will randomly pick any color.

Mitesh Rathod 1:48:04
Good.

Tarun Jain 1:48:06
This is mainly used for correlation. So once you display any correlation rate and if you want to check which two entries are similar during that time, you can check the points of it.

Mitesh Rathod 1:48:18
What is?

Tarun Jain 1:48:20
Uh, yeah.

Mitesh Rathod 1:48:23
No, no. Thank you.

Tarun Jain 1:48:25
OK.

Tirth 1:48:25
So in such in such graphs of uh, you know the scatter plot, there are some, you know, uh lines that we can draw.

Tarun Jain 1:48:36
4 Scatterplot.

Tirth 1:48:36

To see, yeah, you know you mentioned it into like you give the example of elephant and the king and the queen, so animals and then so you know we do some grouping and then there is a line which will show how far away they are from each other. So just consider this into two.

Mitesh Rathod 1:48:38
OK.

Tarun Jain 1:48:44
Yeah, yeah, yeah, yeah.

Tirth 1:48:53
2D uh graph. This is kind of a scatter plot, right?

Tarun Jain 1:48:56
Yeah, this is a scatter plot.

Tirth 1:48:58
So then you know we can have a line and we can know the distance like how far they are away from something like that. So such things.

Tarun Jain 1:49:05
That won't be possible because here it is randomly picking, not randomly, but if you see these points are different.

Tirth 1:49:14
Mm.

Tarun Jain 1:49:15
Even if you put line, I don't think we can.

Tirth 1:49:18
Like a mean or something of. I don't know, maybe I'm wrong. Maybe I'm just overthinking, OK?

Tarun Jain 1:49:23

No, in scatter plot usually will not have lines, so I'll tell you where this is used.

Uh, there is a algorithm in neural network called linear regression.

Tirth 1:49:35

Mhm.

Tarun Jain 1:49:36

And I'll tell use case. Use case is to predict.

Tirth 1:49:38

Mhm.

Tarun Jain 1:49:42

Uh, what is the best use, Kish?

So let's suppose you're working in some real estate use case.

And you have latitude of some place, you have longitude of that place, you have area in what area that is in and you have other important details and the target variable is to predict the price of that particular property.

OK, So what are the features you have? You have latitude, you have longitude, you have area, and you will also have.

Mitesh Rathod 1:50:20

Landmark.

Tarun Jain 1:50:22

Landmark and X number of features and the target is to predict price. So when we are predicting price, usually linear regression will draw a straight line once it trains anything.

Tirth 1:50:37

Mhm.

Mitesh Rathod 1:50:38

It.

Tarun Jain 1:50:38

Linear regression will draw a straight line.

So when we are displaying this entire straight line, we'll have these entries. This entry should be very close to that line. So during that time we'll get to know which columns are very close to the pricing variable. So I'll just show one example linear regression.

Lord.

So if you see this, there is a single line which is straight and then you have rainfall, umbrella sold. If you see umbrella sold is the column. How much impact as umbrella sold made on rainfall?

Mitesh Rathod 1:51:12

OK.

Tirth 1:51:16

OK.

Hmm.

Hmm.

Tarun Jain 1:51:26

So you have a line at it and then you have scatter plot. If it is very close to that line, right? That means that particular column is very important.

Tirth 1:51:28

Mhm.

Mhm.

I see. I see.

Tarun Jain 1:51:36

And you usually use scatter plot only because scatter plot is to compare between 2 variables and you do this with all the fields. So you will do price against area.

Tirth 1:51:41

Mm.

That is what I was looking for, the linear regression part, yeah.

Tarun Jain 1:51:55

Price Vex's landmark.

You will do price, vexes, other fields that you have. You will do it with all the fields.

Tirth 1:52:04

Um.

Tarun Jain 1:52:05

And then you will say, did that particular column impacted my model training or not? If it impacted, then fine. If it didn't impact, I will remove that column. I will train the model again.

Tirth 1:52:11

Hmm.

Hmm.

Tarun Jain 1:52:20

So usually scatter plot is used here in linear regression or in simple words whenever you're comparing 2 fields.

Mitesh Rathod 1:52:23

OK.

Tarun Jain 1:52:28

And the best example for that is correlation.

This is also scatterplot only.

So subplot this particular value is something that I wanted to talk about. So you have two fields. One is same example you have two input variable. If you look at this value 2 comma one comma one that means.

I need to have total 2 rows which is 2 rows.

And two rows one column. So if you see two comma one is nothing but two rows one column. The first one whatever I have, I have to display it in this format only bar plot and in this second which is 2 comma one.

Tirth 1:53:09
Mhm.

Mitesh Rathod 1:53:15
This thing.

Tarun Jain 1:53:20
If you see I have total 2 rows, one column, the second entry, whatever I have, I have to use the colors that I have of my own. So here I just wanted to show how subplot works if I do 1 comma 2.
One comma two it will change instead of having two rows. Now it is 2 columns.
So what is the second? If you define two, that means whatever changes you make here will be reflected in this particular chart.

Mitesh Rathod 1:53:43
OK, good.

Tarun Jain 1:53:56
So this two is nothing but the second graph.

Tirth 1:53:59
OK.

Tarun Jain 1:53:59
So I have one row, two column. The second column is the this particular what you call approach. So instead of this what I'll do is.
I will make it as IST and I will only display episode.

Mitesh Rathod 1:54:22
This.

Tarun Jain 1:54:25

And now I can make it to 1.

Is this clear? Subplot is nothing but let's suppose you are displaying two or three graphs at once. If I do 3 graphs, what are the different combinations I can try? I want to display 3 graphs.

Mitesh Rathod 1:54:32

Yes, it's.

Then.

Yeah.

Tirth 1:54:50

You can.

Tarun Jain 1:54:52

Manis.

Tirth 1:54:54

At the bar plot you have have the.

Tarun Jain 1:54:56

3 comma one or it can be 1 comma three. There is no other combination, but if I have four I have.

Tirth 1:55:01

Um.

Tarun Jain 1:55:05

Two comma 2.

Tirth 1:55:08

B comma one.

Tarun Jain 1:55:08
I have four, I have four comma one. So now here how much time can I repeat it?

Mitesh Rathod 1:55:10
Yeah.

Tarun Jain 1:55:18
So this will be one. So in my two. So let's suppose I have two.
Two rows, 2 columns, whichever is at this position which is 2 cross 2.

Mitesh Rathod 1:55:25
OK.

Tarun Jain 1:55:30
Wait, let me show an example.

Mitesh Rathod 1:55:37
Mhm.

Tarun Jain 1:55:39
It.

Mitesh Rathod 1:55:44
OK.

Tarun Jain 1:55:44
So you understood right? You have to close to. The first plot is whatever the syntax is. Then the second plot is this particular syntax. Third plot is in the sense it will start from.

Tirth 1:55:47
Mhm.

Tarun Jain 1:55:57

The 2nd row, first column and the 2nd row, second column is 4th, so it will go like 1-2, then three and then 4.

Tirth 1:56:05
Yeah.

Tarun Jain 1:56:09
You understood the syntax on how to use subplot.

Tirth 1:56:13
So I think it will take some time to play around with it.

Tarun Jain 1:56:16
Uh, like beside many of comments, but the thing with.

Mitesh Rathod 1:56:17
No, if you if you.

Tarun Jain 1:56:22
Yeah.

Mitesh Rathod 1:56:23
If you just hold the like a subplot the function, it will give us a function doc. You can see here like end rows, end columns, index. OK so we can plot like.

Tarun Jain 1:56:32
Huh. This is your.
Yeah.

Mitesh Rathod 1:56:41
Yeah, so like the two by two comma two is like 2 rows, 2 columns and one is like a first index, then like a 222. So it's a it's placed on the second index. So we can consider it like a yeah, correct.

Tarun Jain 1:56:48

Correct.

Which is this one which will come here?

Then two cross 2-3 is nothing, but this is the third index and this is the 4th index.

So you for two columns, you're just defining your shape.

Mitesh Rathod 1:57:03

But they have like 4 blocking.

Hey.

Tarun Jain 1:57:08

Third column is where you want to place your particular plot.

Mitesh Rathod 1:57:13

Position we can scale for.

Tarun Jain 1:57:14

Ha position.

You can say placement or position or index.

But the thing with matplotlib is the syntax. If you notice everywhere I'm just using PLT dot PLT dot PLT. I'm not using any variables. So the only thing what we need to understand is if I use scatter, what are the input variables I need to give?

Mitesh Rathod 1:57:34

Objects, yeah.

Tarun Jain 1:57:40

Color marker label will come everywhere. These are common syntax in scatter. It's common syntax in bar plot. It's also common syntax in line plot. Usually I don't think bar plot has marker, but color is there.

Here you also if you see PLT dot bar then PLT dot X label, X label is distinct. If you add Y label it will add below. The only thing is we need some practice. This is where we usually perform EDL.

Yeah, this is it from matplotlib to just to summarize, we had line.

Then we had bar plot, then we had histogram, then scattered and then a plot.

Mitesh Rathod 1:58:28

Please.

It's still there.

Bye.

Tarun Jain 1:58:36

So in Pi if you use explode.

It will add that outside.

And if you use auto PCT it will add label.

On the diagram.

So this label is nothing but a percentage distribution.

And whatever repo you have here, right? So far, where are we? The 100 days of code?

Right.

If you see a numpy one day we spent pandas 2 days, then matplotlib just one day.

But if you look at EDA, as soon as you complete numpy pandas matplotlib, definitely we will not know where to use it because most of the time we'll forget the syntax.

So this is where EDA is very important, where I've specifically added two days after even adding two days if you look at mini projects.

In many projects you'll find here is one EDA example, then rain prediction. If you notice in rain prediction also you'll have EDA, so you will never build any model without any EDA. So EDA is first process and then model building.

So this is what eat map looks like. So we know the syntax data dot is null.

What does this do? This will just print if there is any empty data or not. So whatever black you see, there is no empty data over there. Wherever you see white, that means there is empty data.

So sometimes we use heat map, but the best logic is data dot T's null dot sum.

Plot is a fancy thing that you're doing. Yeah, what you call ED.

Now you also know the syntax. What is this? Can anyone tell me?

Tirth 2:00:50

Correlation.

Tarun Jain 2:00:51

Hello correlation and then just add it map. So then once you add it map if you look at your diagonal it is always one and then you can check this color wherever you have grey color 0.9 that means it's correlated. So once your data increases you will not know the values it's like.

Mitesh Rathod 2:00:52
Oh yes.

Tirth 2:00:54
Hmm.

Mitesh Rathod 2:00:54
OK.
We.

Tirth 2:00:59
Mm.

Tarun Jain 2:01:11
It will have to manually go and check it. So if I directly add heat map and if I check the color, if it is one, I have to check with what is this color? Cream color. So wherever cream color is there, that means it's highly correlated.

Tirth 2:01:24
Yeah.

Tarun Jain 2:01:29
This is what we usually do in EDA, just plot.
So there are three to four examples in this particular repo, but we'll try to cover one or two of them. Since we have weekend Saturday, Sunday and you know most of the syntax of numpy, pandas and matplotlib, what you can do is you can directly go and explore.
So if you see I'm loading VG sales which is the video game sales data, I have name, I have platform here then I have global sales. Now what I need to do is I want to filter the publisher as Nintendo and I want to check what is the US UA sales that is more

than 10.
And I want to know the name of the video game.
So for the publisher Nintendo, I want to know what are the sales in Europe which has more than five or sales in Japan which are more than three and then get the name of those. So those are filtering techniques that you will usually do in.
ED.
So if you look at all the commands in the code, we have already covered those.
Probably the only new thing that you will see is we are using C1 in some of the places.

Mitesh Rathod 2:02:46
OK.

Tarun Jain 2:02:52
See what is similar to Matplotlib.

Tirth 2:02:56
OK.

Tarun Jain 2:02:57
So here if you see C1 dot SB.
I'm just using SB dot count plot, so the syntax is same compared to what you use in pandas. I mean matplotlib.
So I'll request spend 2 days of time in playing around with EDA. This is not helpful when it comes to your NLP space, right? In NLP if you're playing around with text usually will not come across this, but if you build any use use cases.
That as sales agent or data analysis agent.
During that time you will have to build your custom tools.
So for custom tools, pandas and matplotlib will come back. Either you can go with pandas or matplotlib or your custom tool will be SQL, which is an alternative of.
On this.
Yes.
Even SQL will just give you command. If you want to display the graph for it, matplotlib will again come here.
Since we have a new case related to sales agent, having understanding of pandas

and matplotlib is very useful.

Is this clear?

Tirth 2:04:25

So far, yes.

Mitesh Rathod 2:04:26

Yes.

Tarun Jain 2:04:27

Obviously you will need practice even if you look at pie chart rate. When I showed pie chart I forgot what explode was.

Tirth 2:04:29

Yeah.

Oh.

Mm.

Tarun Jain 2:04:35

So these are something you will remember only PLT dot pipe explored auto PCT. No one can buy this, so it comes only when you see documentation or some examples.

Tirth 2:04:40

Yeah.

Tarun Jain 2:04:52

So what you can try to do is you have this particular URL, the same repo that we are using. In notebooks you have EDA. Start with pokémon EDA because this is very detailed. I have experimented with so much of commands if you face.

Mitesh Rathod 2:04:52

OK.

Tarun Jain 2:05:09

Any issues with any comment you can let me know if you see this is very detailed.

There are so many plots here.
And then you can also check with video game sales. The only thing what you're doing in EDA is you're just playing around with data, trying to understand certain pattern.
This is something that will confuse, but we'll cover that on Monday.

Tirth 2:05:42
OK.

Tarun Jain 2:05:42
But The thing is, if you get some time on weekends, just try to play around with EDA. It's very fun project. I took pokémon and video game sales because it was something that was fun. If you in case you have any interest of your own, let's suppose you want to explore something related to food.
And you want to play around with food data, just come to Kaggle, type your interest food.
Click on Datasets. Pick any data set, not image.

Tirth 2:06:07
Mitesh.

Mitesh Rathod 2:06:09
It.

Tarun Jain 2:06:14
If you Scroll down here if you see you have some entries and then if you click on code just use filter as most vote.
Most of the time you will find EDA here.

Tirth 2:06:25
OK.
OK.

Tarun Jain 2:06:29
If I click on video game sales.

So if you see the first thing is EDA in Kaggle, most of the code that you will find will be EDA. If you see EDA video game sales, EDA video game sales, then video game sales, EDA, EDA video game sales, EDA is common.

Mitesh Rathod 2:06:37

Yeah.

Tirth 2:06:38

M.

Mm.

Hmm.

Tarun Jain 2:06:51

Every single data and it starts with EDA and they spend most time in EDA only. So you'll find at least 1,00,000 notebooks on Kaggle which is related to EDA.

So just have to check your interest. If it is sales related, just type sales, click on data set, get the data set if I click on coffee sales.

Go to code.

Click on most votes. The first thing is Ed.

Mitesh Rathod 2:07:19

It's.

Tarun Jain 2:07:24

So you will just have to download the data. Once you download the data, start with pandas, use the commands you already know, then try experimenting with new functions.

And in order to experiment, click on code and make sure you select most votes so that you'll get to know which notebooks are good enough.

So you can try with this coffee sales as well if anyone has interest in coffee.

Mitesh Rathod 2:07:54

Of course.

Tarun Jain 2:07:57

Is this clear? Just make sure you pick any data of your choice. Just go to Kaggle, type your interest, get the data set and play around. And if you want to know the syntax, you can check this pokémon EDA and video game sales first.
And then try with new data. So this will be your assignment. You'll have to pick any data set of your choice and then perform ED.

Tirth 2:08:23
Mhm.

Tarun Jain 2:08:24
So whatever syntax is there, it's similar to what we have already seen, but the only thing is experimentation is required. You can take one week of time, probably next Sunday or next Friday we will check what EDA you have done.
I didn't test what you call. I mean, I didn't review your work piece, but I will definitely review ADA.
Instead of given that, I'll make.
Uh, EDA on your.
Interested data and the syntax is same. Just come to Kaggle, search for any name, then data set and once you click on data set you just have to click on download. That's it.

Tirth 2:09:17
Mhm.

Tarun Jain 2:09:19
Download a zip file.
And now if I click on this thing you have Walmart sales dot CSE. This CSE file you need to upload.

Mitesh Rathod 2:09:32
Nothing.

Tarun Jain 2:09:40
So we are done with the pending topic that we had. So Monday we'll directly jump

into Langchain.
So now the pending topics are rag, agents and fine tuning, the three major topics.

Mitesh Rathod 2:09:56
OK.

Tarun Jain 2:09:58
But make sure you spend some time in EDA this weekend. Also you can spend next weekend. Also if you have time you can spend because the more you experiment with data you will get to know certain patterns.

Tirth 2:10:14
That's good.

Tarun Jain 2:10:19
Uh, anything else?

Mitesh Rathod 2:10:21
No, I'm good.

Tarun Jain 2:10:24
You can also take stock price. If in case you have interest in stock, you can just search for stock in Kaggle, get the data set and play around with it.

Tirth 2:10:39
Sounds good.

Tarun Jain 2:10:43
OK, if you have any questions in ETA, right? Most of you guys already have my e-mail. If not, you can also message me on WhatsApp. I'll reply very quickly.

Mitesh Rathod 2:10:44
OK.

Tirth 2:10:51
OK. OK. Thank you so much, David.

Mitesh Rathod 2:10:53
OK.

Tarun Jain 2:10:54
Yeah.

Mitesh Rathod 2:10:55
Thanks. Bye.

Tirth 2:10:55
Thank you. Bye.

◉ **Margi Varmora** stopped transcription