Topic	Correlation	
Class Description	**Students get introduced to the concept of "correlation" and how to identify if two data sets are correlated visually by drawing plots. Students also learn how to calculate correlation and write a python program to calculate correlation.**	
Class	**C106**	
Class time	**45 mins**	
Goal	• Understand the concept of correlation • Identify if two data sets are correlated using visual charts • Calculate correlation and write a python program for it	

Resources Required	• Teacher Resources
	  ○ Google Colab
	  ○ Laptop with internet connectivity
	  ○ Earphones with mic
	  ○ Notebook and pen
	• Student Resources
	  ○ Google Colab
	  ○ Laptop with internet connectivity
	  ○ Earphones with mic
	  ○ Notebook and pen	

Class structure	**Warm Up** **Teacher-led Activity** **Student-led Activity** **Wrap up**	**5 mins** **15 min** **15 min** **5 min**

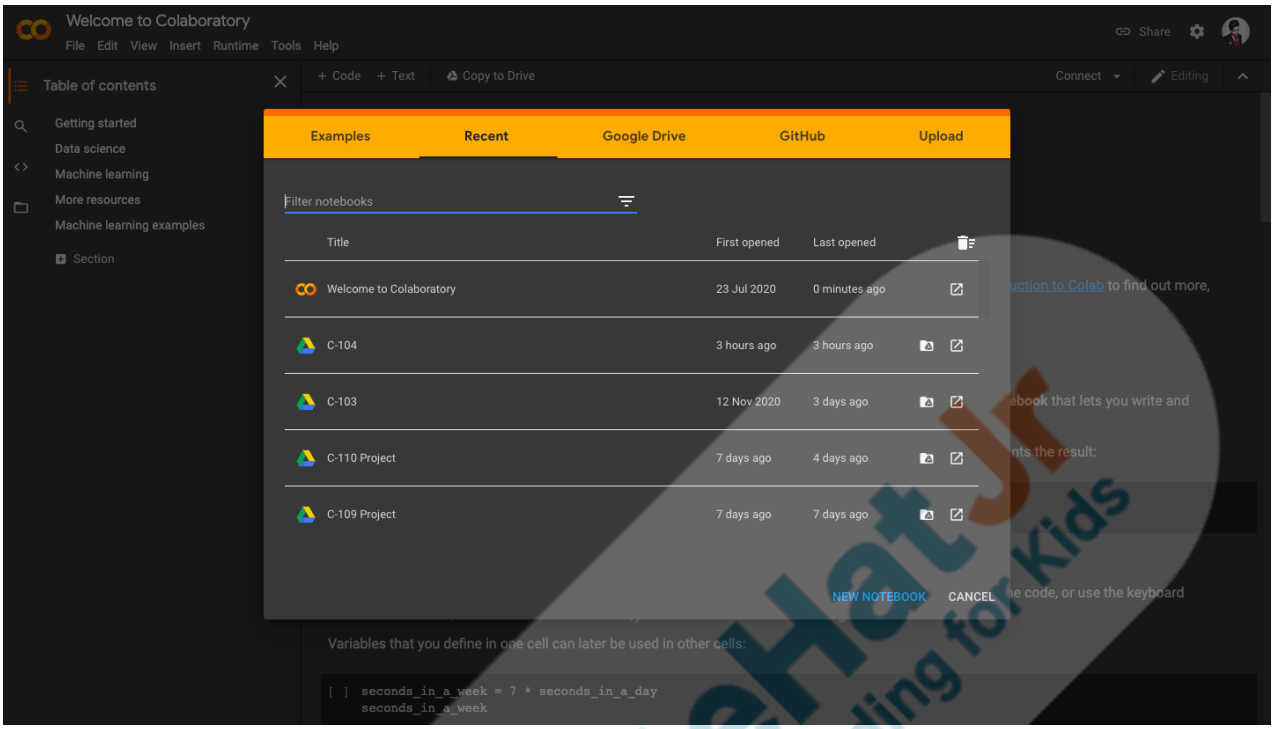
CONTEXT

- **Review central tendency and standard deviation**

Class Steps	Teacher Action	Student Action

Step 1: **Warm Up** **(5 mins)**	Hello! We've been learning statistics and python programming since the last few classes. Do you remember what we have covered in the last few classes?	**ESR:** - We learned how to calculate the central tendency of data. - mean, median and mode. We also learned how to calculate standard deviation in data.
	Can you describe in your own words - what is the central tendency of data?	Central Tendency of data is a value which tries to describe the central position (where most of data is centred) of data. There are different ways in which central tendency of data can be calculated. Mean, median and mode are different methods through which we can calculate the central tendency of data.
	Awesome. How would you describe "standard deviation"?	**ESR:** Standard deviation is a measure of how much the members of a data set differ from the mean.
	Great! So far, we have been restricted to one data set. In this class, we will explore more than one data set and learn to identify if the two data sets are related to each other.	-
Teacher Initiates Screen Share		

CHALLENGE • **Look at data containing record of temperature for each day and sales of ice cream and cold-drinks from a public store** • **Visually identify if the data sets are correlated** • **Calculate the correlation index on a spreadsheet**		
Step 2: Teacher-led Activity (15 min)	We will be using Google Colab for this class! *<Teacher opens a new Google Colab>* *<Watch the short introduction video about Colab if the child has not worked with Google Colab before>* *<Teacher opens the link from **Teacher activity 2** and watch the video>* To open a new google colab, refer to ***Teacher activity 3.***	
	In Colab every project is called a **notebook**. When we open a Colab we see a pop up where we can select our previous notebook to continue our work or create a new notebook to work on a new project. We'll create a new notebook. Here we can write python code as well as text.	



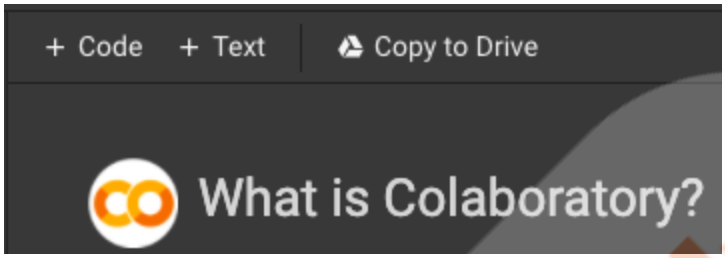
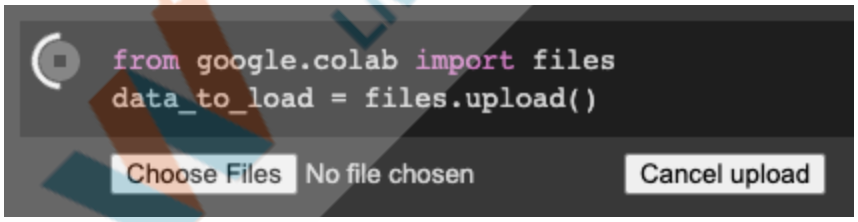
Can you guess how we can write code and text?

Yes, To write code we click on the code button. A code cell opens up where you can write your code and press the run button to execute your code.

<The teacher clicks on the code button and types print("hello world") in the code cell and clicks on the run button>

Same way we can add the text in the notebook. Text can be used for general purpose like:

-Adding a heading.

	-Adding an explanation on what your code block is doing. -Adding instructions.	
		
	Uploading and importing files in Colab is also very easy. To upload the files in Colab we just have to write a small piece of code. *<Teacher writes the following code in code cell>* Code:- **from google.colab import files** **data_to_load = files.upload()** a choose file button will appear. by clicking on the button we can upload the files from our local system.	
		
	Have you ever wondered if two different events can have some relation to each other? For example - can the number of car accidents in a day and temperature of a particular day be related?	**ESR:** They might have some relation. People might get more frustrated when the temperature is high. It might lead to more road rage and more accidents!

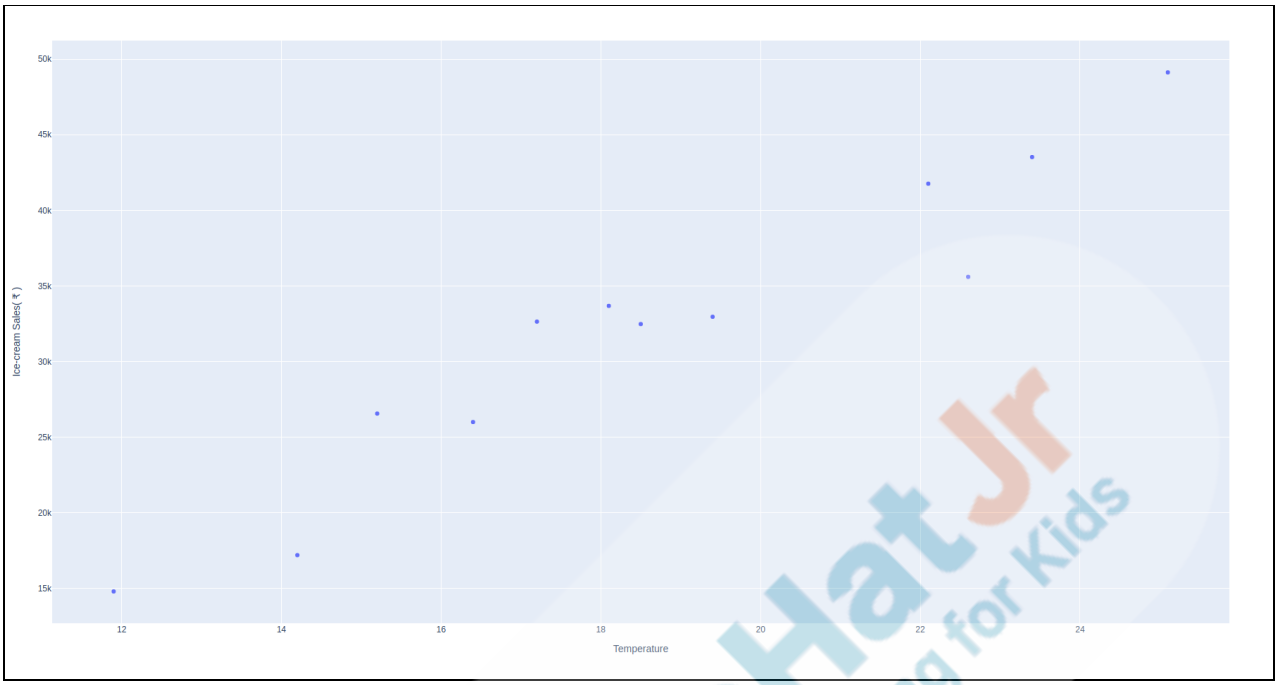
	What about the temperature of a day and sale of ice-cream in a store?	**ESR:** High temperature might lead to high sales of ice-cream in the store.
	Let's check one such data. We have some data where temperature of each day is recorded. It also contains data of sales of ice-cream in a public store. Let's plot this data as a scatter plot and see how it looks. Can you help me make the scatter plot? What would be on the X-axis? What would be on the Y-axis?	**ESR:** X- axis can be temperature Y- axis can be the sales from ice-cream
	*<CSV can be found at **Teacher Activity 4**>.* *Teacher writes code to create a scatter plot for Temperature vs Sales from ice-cream.* *Teacher runs the code to show the data visualization.*	*Student guides the teacher in writing the code for creating the scatter plot for Temperature vs Sales from ice-cream.*

```
[1] from google.colab import files
    data_to_load = files.upload()

    Choose Files  Ice-Cream ...ure data.csv
    • Ice-Cream vs Cold-Drink vs Temperature - Ice Cream Sale vs Temperature data.csv(text/csv) - 262 bytes, last modified: 17/12/2020 - 100
    Saving Ice-Cream vs Cold-Drink vs Temperature - Ice Cream Sale vs Temperature data.csv to Ice-Cream vs Cold
```

```
[3] import plotly.express as px
    import csv

    with open("Ice-Cream vs Cold-Drink vs Temperature - Ice Cream Sale vs Temperature data.csv") as csv_file:
          df = csv.DictReader(csv_file)
          fig = px.scatter(df,x="Temperature", y="Ice-cream Sales( ₹ )")
          fig.show()
```



	What do you see? What happens when the temperature increases? What happens when the temperature decreases?	**ESR:** When the temperature increases, the sale of ice-cream goes up. When the temperature reduces, the sale of ice-cream goes down.
	You see the data are not scattered on the graph and are close towards a central line. Such data are said to be highly correlated. Data sets can also be inversely correlated. What do you think that means? Can you think of data sets which might be inversely correlated?	**ESR:** One data set increases when the other data set reduces? **ESR:** Temperature data vs sale of warm clothes in a store.

	Yes! How do you think the scatter plot for such a data set would look like?	**ESR:** varied
	We have a data set for consumption of coffee vs Hours of sleep. How do you think they might be correlated? *<CSV is available at **Teacher Activity 5**>*	**ESR:** When cups of coffee decrease, hours of sleep increase or we can say the two data sets are inversely correlated.
	Let's draw a scatter plot and see. Can you help me draw a scatter plot visualization for the data. *Teacher writes code to draw the scatter plot.*	*The student helps the teacher write the code.*

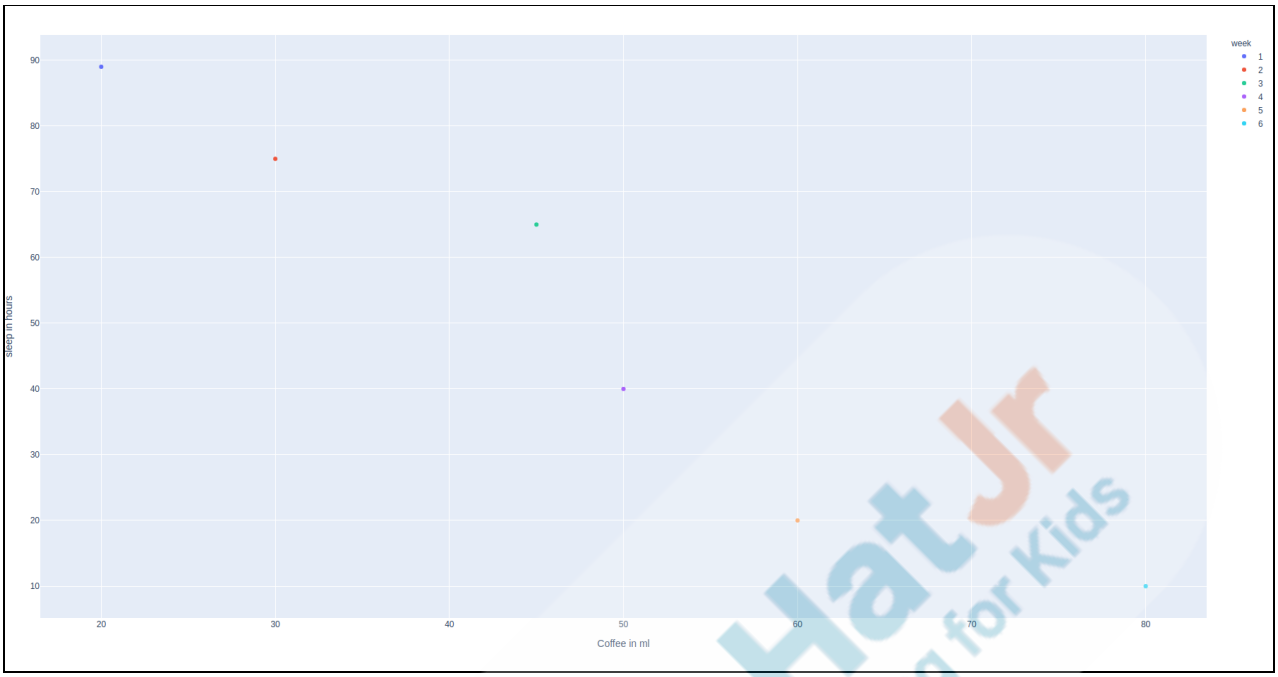
```
[4] from google.colab import files
    data_to_load = files.upload()
```

```
Choose Files  cups of coff...of sleep.csv
  • cups of coffee vs hours of sleep.csv(text/csv) - 81 bytes, last modified: 17/12/2020 - 100% done
Saving cups of coffee vs hours of sleep.csv to cups of coffee vs hours of sleep.csv
```

```
import plotly.express as px
import csv

with open("cups of coffee vs hours of sleep.csv") as csv_file:
      df = csv.DictReader(csv_file)
      fig = px.scatter(df,x="Coffee in ml", y="sleep in hours", color="week")
      fig.show()
```

	Let's run the code to check the output. What do you see?	**ESR:** We see a falling graph where the data is still close to the central straight line.



	Yes! This is how an inverse correlation looks like. We can also calculate correlation value. A correlation of 1 means the two data sets are closely correlated. This will be a rising graph where the data points are close to a central line. A correlation of -1 means that the two data sets are inversely correlated. This will be a falling graph where the data points are close to a central line. A correlation of 0 means that the two data sets are not correlated at all! The data points will be scattered on the graph. Correlation always lies between -1 and 1	*The student asks questions to understand correlation.*

	Let's look at how to calculate correlation coefficient using the corrcoef() function in numpy library. We will calculate the correlation coefficient of temperature and ice-cream sales data. 1. Teacher imports numpy library in code. (Make sure numpy is pre-installed using pip3 install numpy) 2. Convert temperature data and ice-cream sales data into arrays. Make sure that each data set is converted into a float value first. (by default each data set is a string) 3. Use corrcoef() function and pass the two datasets to it. Store the output in a variable. 4. Print the correlation coefficient on the screen. What is the result?	**ESR:** 0.95

```
import plotly.express as px
import csv
import numpy as np

def plotFigure(data_path):
    with open(data_path) as csv_file:
        df = csv.DictReader(csv_file)
        fig = px.scatter(df,x="Temperature", y="Ice-cream Sales( ₹ )")
        fig.show()

def getDataSource(data_path):
    ice_cream_sales = []
    cold_drink_sales = []
    with open(data_path) as csv_file:
        csv_reader = csv.DictReader(csv_file)
        for row in csv_reader:
            ice_cream_sales.append(float(row["Temperature"]))
            cold_drink_sales.append(float(row["Ice-cream Sales( ₹ )"]))

    return {"x" : ice_cream_sales, "y": cold_drink_sales}

def findCorrelation(datasource):
    correlation = np.corrcoef(datasource["x"], datasource["y"])
    print("Correlation between Temperature vs Ice Cream Sales :-  \n--->",correlation[0,1])


data_path  = "Ice-Cream vs Cold-Drink vs Temperature - Ice Cream Sale vs Temperature data.csv"
datasource = getDataSource(data_path)
findCorrelation(datasource)
plotFigure(data_path)
```

```
Correlation between Temperature vs Ice Cream Sales :-
---> 0.9575066230015955
```

	What does this tell?	**ESR:** The two data sets are positively correlated. They also have a high correlation.
	Alright. You now know how to create a visual representation of two data sets and identify if the two data sets are correlated or not. Now, I am going to give you two different data sets. 1. One data set compares the number of hours of TV watched in a week on average vs the size of television.	

<table>
<tr><td></td><td>2. Another data set is of the number of days each student has been present in college in a year vs the percentage of marks scored in the half-yearly exams.

Can you plot the data for each and visually guess if they are correlated? You can then also use the in-built correlation coefficient function to calculate the correlation.</td><td>ESR:
Yes</td></tr>
<tr><td colspan="3">Teacher Stops Screen Share</td></tr>
<tr><td></td><td>Now it's your turn. Please share your screen with me.</td><td></td></tr>
<tr><td colspan="3">• Ask Student to press ESC key to come back to panel
• Guide Student to start Screen Share
• Teacher gets into Fullscreen</td></tr>
<tr><td colspan="3"><u>ACTIVITY</u>
• Student writes python program to calculate correlation index</td></tr>
<tr><td>Step 3:
Student-Led Activity
(15 min)</td><td>What do you think would be the relationship between the number of hours of TV watched in a week on average vs the size of television? How would their correlation be?

<i><CSV is available at Student Activity 2></i></td><td>ESR:
People might watch more TV in a week as the size of television goes up.</td></tr>
</table>

	Let's visually plot the data and check.	*The student downloads the data.* *Student writes code to :* *- read the data* *- use plotly to draw a scatterplot for the data*

```
[8] from google.colab import files
    data_to_load = files.upload()
```

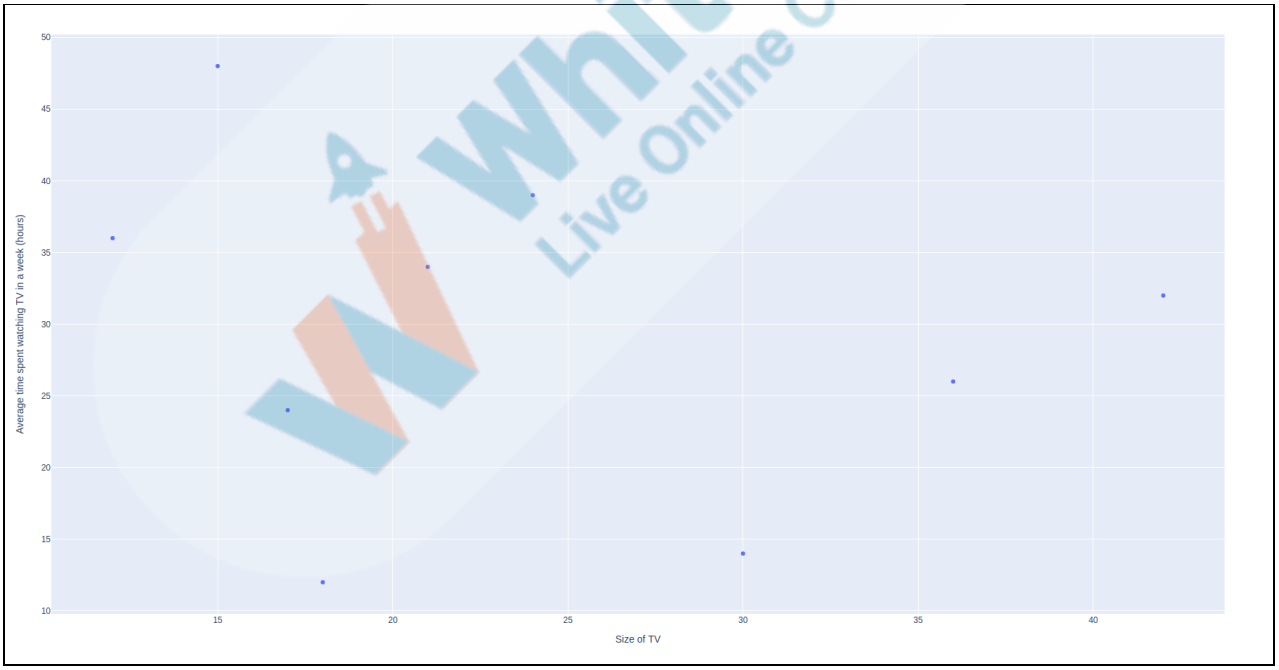
Choose Files | Size of TV, ... (hours).csv

- Size of TV, Average time spent watching TV in a week (hours).csv(text/csv) - 123 bytes, last modified: 17/12/2020 - 100% done

```
Saving Size of TV, Average time spent watching TV in a week (hours).csv to Size of TV, Average time
```

```
import plotly.express as px
import csv

with open("Size of TV, Average time spent watching TV in a week (hours).csv") as csv_file:
      df = csv.DictReader(csv_file)
      fig = px.scatter(df,x="Size of TV", y="\tAverage time spent watching TV in a week (hours)")
      fig.show()
```



	Look at the scatter plot. Are the points close to any central line. How do you think is the correlation between the average number of hours of TV watched vs size of the television.	**ESR:** No! They are scattered. **ESR:** They both are not at all correlated.
	Let's calculate the correlation index and check its value.	*Student writes code to calculate the correlation index and finds it is close to 0.*

```
import csv
import numpy as np


def getDataSource(data_path):
    size_of_tv = []
    Average_time_spent = []
    with open(data_path) as csv_file:
        csv_reader = csv.DictReader(csv_file)
        for row in csv_reader:
            size_of_tv.append(float(row["Size of TV"]))
            Average_time_spent.append(float(row["\tAverage time spent watching TV in a week (hours)"]))

    return {"x" : size_of_tv, "y": Average_time_spent}

def findCorrelation(datasource):
    correlation = np.corrcoef(datasource["x"], datasource["y"])
    print("Correlation between Size of Tv and Average time spent watching Tv in a week :-  \n--->",correlation[0,1])


data_path  = "Size of TV, Average time spent watching TV in a week (hours).csv"
datasource = getDataSource(data_path)
findCorrelation(datasource)
```

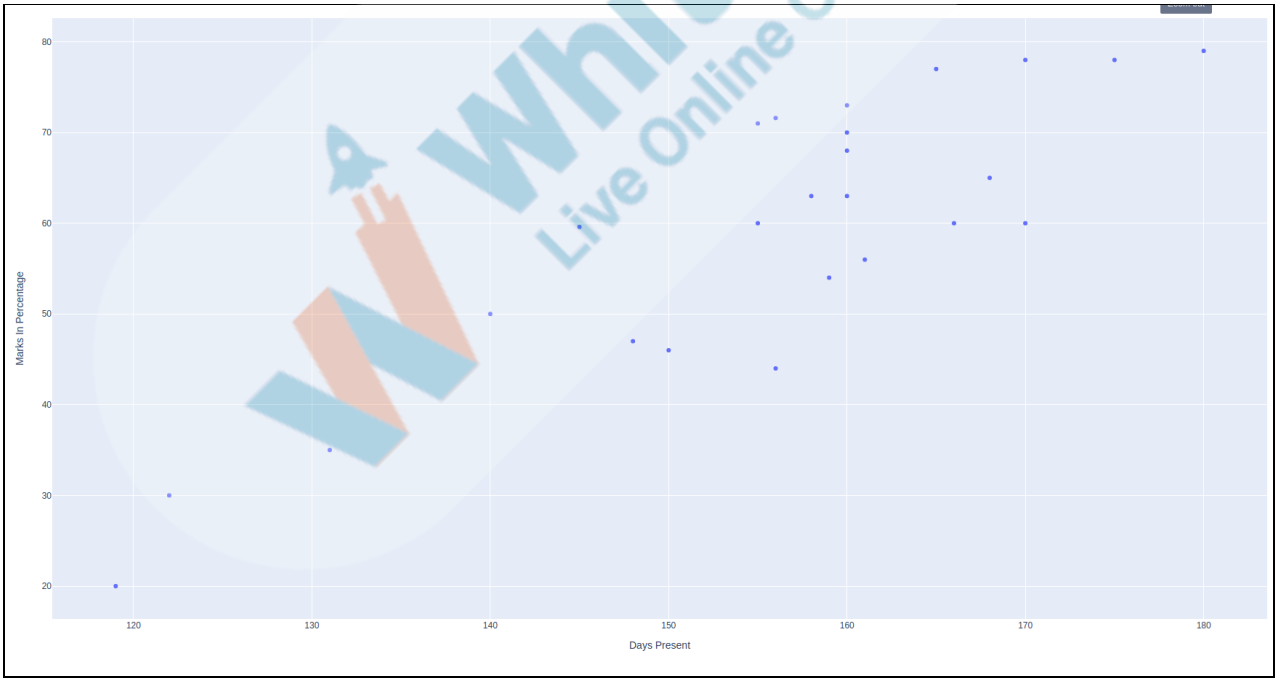
```
Correlation between Size of Tv and Average time spent watching Tv in a week :-
---> -0.21596489617950243
```

	Awesome. Now let's take a look at the next dataset. We have the number of days students attended college vs the marks they scored in their exams. How do you think these two would be correlated?	**ESR:** varied

	*<CSV is available at **Student Activity 3**>*	
	Let's plot them visually and check if your guess is right	*The student downloads the data.* *Student writes code to :* *- read the data* *- use plotly to draw a scatterplot for the data*

```
[15] import plotly.express as px
     import csv

     with open("Student Marks vs Days Present.csv") as csv_file:
           df = csv.DictReader(csv_file)
           fig = px.scatter(df,x="Days Present", y="Marks In Percentage")
           fig.show()
```



	What do you think is the correlation between the number of days one attends the classes vs the percentage of marks scored in the exams?	**ESR:** The two data are positively correlated.
	Let's calculate the correlation index to verify.	*Student writes code to calculate the correlation index and finds it is close to 1.*

```
[16] import csv
     import numpy as np


     def getDataSource(data_path):
         marks_in_percentage = []
         days_present = []
         with open(data_path) as csv_file:
             csv_reader = csv.DictReader(csv_file)
             for row in csv_reader:
                 marks_in_percentage.append(float(row["Marks In Percentage"]))
                 days_present.append(float(row["Days Present"]))

         return {"x" : marks_in_percentage, "y": days_present}

     def findCorrelation(datasource):
         correlation = np.corrcoef(datasource["x"], datasource["y"])
         print("Correlation between Marks in percentage and Days present :-  \n--->",correlation[0,1])

     data_path  = "Student Marks vs Days Present.csv"
     datasource = getDataSource(data_path)
     findCorrelation(datasource)
```

```
Correlation between Marks in percentage and Days present :-
---> 0.86288947614385
```

	Awesome.	-
Teacher Guides Student to Stop Screen Share		
FEEDBACK • **Appreciate the student for their efforts** • **Identify 2 strengths and 1 area of progress for the student**		
Step 4: Wrap-Up (5 min)	Can you capture what we learned in today's class?	**ESR:** - We learned about correlation. - How to plot data and identify if the data sets have some relation be. - We also learned about correlation coefficient and how to calculate it using python program.
	Correlation is very important in data science. For example, a lot of data scientists spend a number of hours trying to identify data sets to which stock prices might be correlated. You can collect datasets and try to identify if any two datasets are correlated. Correlation is also an important concept in machine learning models.	-

	It can help us predict future data based on the previously collected data. We'll be learning more about it in the coming classes.	
Project Overview	## Correlation **Goal of the Project:** In this project you will have to write the program to find the correlation of the given data sets and plot it on a graph. **Story:** In this world everything is connected or dependent on other things. For example as Global warming increases the ice in the poles start to melt. So, Global warming and ice are related. Based on this, your teacher has assigned you a task of writing an article about this phenomena for your school magazine. To help you on this article, your teacher has given you some data of Student marks and days the student was present. And data for cups of coffee and the amount of sleep. Work on these data to see how they are correlated and dependent on each other in order to present your findings in the article.	

	Write a program to find the e correlation between the given datasets . I am very excited to see your project solution and I know you will do really well. Bye Bye!	
Teacher Clicks 		
Additional Activities	*Encourage the student to write reflection notes in their reflection journal using markdown.* Use these as guiding questions: • What happened today? - Describe what happened - Code I wrote • How did I feel after the class? • What have I learned about programming and developing games? • What aspects of the class helped me? What did I find difficult?	*The student uses the markdown editor to write her/his reflection in a reflection journal.*

Activity	Activity Name	Links
Teacher Activity 1	Solution	https://colab.research.google.com/

		drive/1wVlNWLZsEWb-dUF4qC0IIYqy-Fw7ojY_?usp=sharing
Teacher Activity 2	Colab Introduction	https://youtu.be/inN8seMm7UI
Teacher Activity 3	Google Colab Link	https://colab.research.google.com/
Teacher Activity 4	Ice-Cream vs Cold-Drink vs Temperature - Ice Cream Sale vs Temperature data	https://raw.githubusercontent.com/whitehatjr/correlation/master/data/Ice-Cream%20vs%20Cold-Drink%20vs%20Temperature%20-%20Ice%20Cream%20Sale%20vs%20Temperature%20data.csv
Teacher Activity 5	cups of coffee vs hours of sleep	https://raw.githubusercontent.com/whitehatjr/correlation/master/data/cups%20of%20coffee%20vs%20hours%20of%20sleep.csv
Student Activity 1	Ice-Cream vs Cold-Drink vs Temperature - Ice Cream Sale vs Temperature data	https://raw.githubusercontent.com/whitehatjr/correlation/master/data/Ice-Cream%20vs%20Cold-Drink%20vs%20Temperature%20-%20Ice%20Cream%20Sale%20vs%20Temperature%20data.csv
Student Activity 2	Size of TV, Average time spent watching TV in a week (hours)	https://raw.githubusercontent.com/whitehatjr/correlation/master/data/Size%20of%20TV%2C%09Average%20time%20spent%20watching%20TV%20in%20a%20week%20(hours).csv
Student Activity 3	Student Marks vs Days Present	https://raw.githubusercontent.com/whitehatjr/correlation/master/data/Student%20Marks%20vs%20Days%20Present.csv

Student Activity 4	cups of coffee vs hours of sleep	https://raw.githubusercontent.com/whitehatjr/correlation/master/data/cups%20of%20coffee%20vs%20hours%20of%20sleep.csv