

Start

# Data Science

## Day 1

classmate

Data Science :- data gathering, analysis & decision-making  
Finding patterns in data through analysis and make future predictions.

- Better decisions
- Predictive analysis
- Pattern discoveries

### Data Scientist

- Expertise in
- Machine Learning
- Statistics
- Programming (Python or R)
- Mathematics
- Databases

### Work flow a Data Scientist works

1. Ask the right questions
2. Explore and collect data
3. Extract the data
4. Clean the data
5. Find and replace missing values
6. Normalize data
7. Analyze data, find patterns & make future predictions
8. Represent the result

## Data

Types : • Unstructured - not organized, must be organized <sup>analysis</sup> for "

• Structured - organized, easy to work with

Array, database table to structure or present data

### • Array in Python

Example :

```
Array = [80, 85, 90, 95, 100, 105, 110, 115, 120, 125]
```

```
print(array)
```

## Database Table

- Table with structured data
- consists of rows and columns

Rows - horizontal

Column - vertical

## Variables

- Something that can be measured or counted.
- examples: characters, numbers or time.

# Day 1

classmate

Date \_\_\_\_\_

Page \_\_\_\_\_

## Python

- programming language <sup>widely</sup> used by Data Scientists
- in-built mathematical libraries and functions

## Python Libraries

- Pandas - structured data operations, like import CSV files, create dataframes, data preparation
- Numpy - Mathematical library, has a powerful N-dimensional array object, linear algebra, fourier transform, etc
- Matplotlib - used for visualization of data
- SciPy - linear algebra modules

- Create DataFrame with Pandas
- A data frame is a structured representation of data.

Example:

```
import pandas as pd
d = {'col1': [1, 2, 3, 4, 7], 'col2': [4, 5, 6, 9, 5], 'col3': [7, 8, 12, 1, 11]}
df = pd.DataFrame(data=d)
print(df)
```

Explanation

- Import the Pandas library as pd
- Define data with column and rows in variable d
- Create data frame using the function pd.DataFrame()
- The data frame contains 3 rows, 5 columns
- Print data frame output with the print() function.

We write `pd.` in front of `DataFrame()` to let Python know that we want to activate `DataFrame()` function from Pandas library.

To count columns and rows using Python:

```
count_column = df.shape[1]
```

```
print(count_column)
```

```
count_row = df.shape[0]
```

```
print(count_row)
```