|  |  |
| --- | --- |
| **Session 49 - Introduction to Data Science - Lesson plan** | |
| Class Objective | At the conclusion of this session, students will be able to **list down** 5 steps and **explain** the Obtain and Scrub step in Data Science Process. Also, students will become **familiar** with terminology of Data Science and will **perform** a high-level overview of the dataset. |
| Class Outcomes | Students will be able to -   1. **List** down the 5 stages of Data Science Process 2. **Explain** the Obtain and Scrub step from the Data Science Process 3. **Familiarize** withthe terminology of Data Science - Table, Rows, Columns, Dataset, Data Analysis, Data Science, Numerical and Categorical variables 4. **Name** different tools, programming languages and libraries used for Data Science/Analysis 5. **Perform** high-level overview of data using functions available in Pandas Library |
| Class Time | 60 Mins |
| Teachers Resources | * Laptop or a Desktop with working Camera and a mic * Good Internet Connection * Headset with a mic * Google Account and Google Drive * Google Chrome Browser * Access to Google Colab |
| Student Resources | * Laptop or a Desktop with working Camera and a mic * Good Internet Connection * Headset with a mic * Notepad and a pencil * Google Account and Google Drive * Google Chrome Browser * Access to Google Colab |
| Class Structure | 5 mins : Warm up with the kid  50 mins : Class Activity  5 mins : Conclusion and doubt clearance |

|  |  |  |  |
| --- | --- | --- | --- |
| **Warm Up** | | | |
| 5 mins | A smile on the face | Hello [student’s name]  How are you?  I am great; thanks for asking. How was your day?  Do you remember what we did in the AI Classes?  Could you quickly summarize it for me?   * Revised concepts from Artificial Intelligence.   Brilliant!!  If the student says no, please help them recollect the programs done in the previous class.  Also clarify the doubts (if any) , on the concepts covered till last class. | Hello  I am good, how are you?  Let the student reply.  Student replies whatever she/he could recall from the previous class. |

|  |  |  |
| --- | --- | --- |
|  | **Teacher Activity** | **Student Response** |
| 25 Minutes | **Part 1: Introduction to Data Science**  Who is the newly elected President of the United States of America?  If the student responds with the name of Mr. Joe Biden, then reward the student.  Otherwise, correct the answer as Mr. Joe Biden being newly elected President of the United States of America.  Well-done, can you tell me how did you come to know about this event?  Very Good! We all have access to the information via newspapers, news channels, twitter, books, encyclopedia etc.  This entire process of collecting the data, processing the data, and finding meaningful analysis from it is called the Data Science Process.  It has five steps -  Obtain the Data  Scrub the Data  Explore the Data  Model the Data  Interpret the Data    Today we will cover first two steps of Data Science Project  **Let us understand the steps in detail -**  What do you mean by the word obtain?  Incredibly good, acquiring the information is the first step.  **1. Obtain Data**  The very first step of a data science project is straightforward. We obtain the data that we need from available data sources.  The way you acquired information about who will be the newly elected president of the United States of America, similar way we can obtain the information or data from various sources, such as books, internet, social networking channels etc.  Once we obtain the data, we need to make use of it.  But what do you think, can we use the information available as it is?  Let us take an example of how gold is mined from the goldmines and is made available to everyone in the form of ornaments.    This is the image of Gold Ore Block extracted from mine.  Its weight is approximately 860 Kg.  Can you think of the amount of Gold that can be extracted from this gold ore?  ***If the student replies with weight of the gold extracted up to 1 Kg appreciate them that they are thinking on the correct line but still need some correction.***  As per Wikipedia, from an 860 Kg of Gold Ore block, only 30 grams of Gold can be extracted.  That is a huge difference right.  This shows, not everything that we obtain from the source can be used directly.  To find out meaningful analysis from the information we need to perform some cleaning operation.  **We call it Scrub the Data**  **2. Scrub Data**  If the data is unfiltered and irrelevant, the results of the analysis will not mean anything.  This process is for us to “clean” and to filter the data.  Obtaining the Data from source and Scrubbing it or cleaning it for further processing are two most important steps of any Data Science Project.  **==================================================**  **Part 2: Terminology of Data Science**  **Before we actually work on a data science project, we need to understand the key terminologies of a Data Science.**  **Scenario 1: Can you predict the survival of passengers aboard Titanic Boat when it sank?**  Ask the students if they have heard about the Story of the Titanic Ship.  If the student knows the story, ask him/her to narrate the story.  You can add the details such as no of passengers what exactly happened to the ship.  Narrate the story to student  **Note -** Do not read the story, explain the story in your own words.  Royal Mail Steamer Titanic was a British passenger liner that sank in the North Atlantic Ocean in the early hours of 15 April 1912, after it collided with an iceberg during its maiden voyage from Southampton to New York City. With 1,500 lives lost out of 2,224, it was one of the deadliest commercial peacetime maritime disasters in modern history. The RMS Titanic was the largest ship afloat at the time it entered service and was the second of three Olympic-class ocean liners operated by the White Star Line. The Titanic was built by the Harland and Wolff shipyard in Belfast. Thomas Andrews, her architect, also died in the disaster.  **Let us find out more about what happened to the passengers aboard on the Titanic.**    **Dataset –**  **Dataset** is a collection of data.  This data can be in the form of numbers or characters, images, audio, videos or sometimes this data can have two or more types such as image and text are combined etc.  It is also represented in a tabular format.  **Table –**  A table is way of presenting facts in the form on rows and columns  We have information of passengers travelling by RMS Titanic in the form of rows and columns arranged as a table.  **Row –**  A row is a horizontal group of values within a [table](https://techterms.com/definition/table). It contains values for multiple fields, which are defined by [columns](https://techterms.com/definition/column).  It is also considered as a **record**.  In our table we have record of more than 100s of passengers  **Column –**  Column is nothing but a vertical group of values within a table.  It represents one **attribute** for multiple rows.  Such as **PassengerID, Survived, PClass etc.**  These are also called as attributes or variables  Dataset, Table, Rows, Column, attributes are the most important terms used in a Data Science Projects.  Now, let us have a look at our Titanic Dataset and the information available for analysis.  **What information do we have in the Titanic Dataset?**   * **PassengerId**: unique id number to each passenger * **Survived**: If the passenger has survived then it is mentioned as 1, otherwise 0. * **Pclass**: passenger class * **Name**: name * **Sex**: gender of passenger * **Age**: age of passenger * **SibSp**: number of siblings/spouses * **Parch**: number of parents/children * **Ticket**: ticket number * **Fare**: amount of money spent on ticket * **Cabin**: cabin category * **Embarked**: port where passenger embarked (C = Cherbourg, Q = Queenstown, S = Southampton)   **===================================================**  **Part 3: Various functions from Pandas**  **In the next three sessions we will try to explore the dataset TITANIC Ship and will try to find out who survived in the disaster and present our findings as a blog post or small piece of article.**  **Here is the list of libraries we are using:**   1. [NumPy](https://docs.scipy.org/doc/numpy-1.13.0/user/whatisnumpy.html) : Multidimensional Array and Matrix Representation Library 2. [Pandas](https://towardsdatascience.com/a-quick-introduction-to-the-pandas-python-library-f1b678f34673) : Python Data Analysis Library for Data Frame, CSV File I/O 3. [Matplotlib](https://matplotlib.org/3.1.0/index.html) : Data Visualization Library 4. [Seaborn](https://seaborn.pydata.org/) : Data Visualization Library built on top of Matplotlib. This gives you a cleaner visualization and easier interface to call.   **Pandas functions used in this session -**   |  |  | | --- | --- | | import pandas as pd | pd is used throughout the project to represent the data. | | DataFrame | structure offered by Pandas Library to read the data from the files | | read\_csv() | load the data from the source.  This function helps us read the data from comma separated values files  We need to mention the correct path to the file | | head() | function shows the rows from beginning of the dataset | | tail() | function shows the rows at the end | | info() | function gives the information about each column such as how many entries are there, data type of the attribute etc | | describe() | gives us the description of the dataset using statistical values | | shape | gives us the number of rows and columns available in our dataset |   **Note**  In **head()**, and **tail()**  we need to mention an integer value which will be used to specify how many rows we would like to observe  Here, we are going to observe the first 5 rows and last 5 rows of the dataset. hence we have specified **head(5)** and **tail(5).** | Student replies with Mr. Joe Biden  Student replies,  I came to know via Twitter, Newspaper, News Channels, Parents informed or any other source they have come to know about the election.  Students will respond as yes/no.  Student replies with some weight as per his/her imagination  Student will either response with they have seen the movie, heard the story or have not heard about it |

|  |  |  |
| --- | --- | --- |
|  | **Activity** | **Student Response** |
| 10 mins  Teacher Activity 1  Teacher Activity 2  Teacher Activity 3  Teacher Activity 4  Teacher Activity 5  Teacher Activity 6  Teacher Activity 7  Teacher Activity 8 | Introduction to Data Science  In this session we will understand how to obtain the data from a source and take an overview of it.  Open the StoryofSurvival\_Titanic Voyage Google Colab file from the  <https://drive.google.com/drive/folders/1LgxZIg8THzKKCLCHrqoNn3QR9Y9d2G1c?usp=sharing> <Teacher Drive Link>  Import required Python Libraries    Now mount the google drive content    Follow the below document for steps to mount the drive.  <https://docs.google.com/document/d/12QKURfl_lZ-xf0QjsszinMlhAyTj1TZtyfKwuqWDIBQ/edit?usp=sharing>  **Now, let us load the data.**    **Time to check what is there in our Titanic Dataset.**  **Let us observe first five rows of titanic dataset**    **Now, let us observe last five rows of titanic dataset**    **Let us now check the dimensions of the dataset.**    **Let us take over view of data and find out some information**  Let us summarize our dataset using some statistics.    **Explain the output as various statistical quantities such as count of values, maximum, minimum etc.**  Create a folder on a google drive and name it as either “Story of Survival” or any other name a student feels is good for him/her.  Create a new Google Colab notebook to explore Titanic Dataset  Use this link to create a new Google Colab Notebook <https://colab.research.google.com/> |  |
| Summarize  (5 min) | I hope you enjoyed the class.  Today we have discussed about -   * First two steps of Data Science Project - Obtain and Scrub the Data * Terminology of Data Science - Rows, Columns, Data, Table, Dataset, Numeric and Categorical Variables * How to load, view, describe the data using Pandas? * How to use Pandas to describe the given data in detail? * Different functions to take high level overview of data - head(), tail(), shape, and describe() |  |

|  |
| --- |
| **Practice Questions** |
| 1. Change the name of the titanic\_train\_df to titanic\_dataframe 2. Perform the Data Overview operations on this new data frame 3. Use a different dataset than Titanic Survival Dataset and perform Data Overview. |