**Important links**

Program Associate : Ankita , [ankita.jain@decodr.in](mailto:ankita.jain@decodr.in)  ,   +91 9837036907  
Subject Matter Expert :  Saurabh Singh,  [sourav.singh@decodr.in](mailto:sourav.singh@decodr.in)  
Trainer : John Gabriel  
Trainer LinkedIn Profile link :<https://www.linkedin.com/in/johngabrieltj/>

Meeting Link   <https://global.gotomeeting.com/join/162779637>

**BatchFolderLink**

<https://drive.google.com/drive/folders/1dcSiZH0TDaiUCz41XBE44uQW0kH2mnFe?usp=sharing>

**SyllabusPPT**

<https://drive.google.com/drive/folders/1iuwTUQsotXCwdEjf7d89y0ScNU6toBJL?usp=sharing>

**CaseStudy**

<https://drive.google.com/drive/folders/1JsgHCVqfqAa7mFG5HgFsUmo5HB6eN1jB?usp=sharing>

**Link for Lecture Recordings**

<https://drive.google.com/folderview?id=1hxssWNOoysZAy5rHBFKcnnwwiQUtFJhZ>

**Slack Video**

<https://youtu.be/Z68vLxQWPg0>

**GoToMeeting Video**    <https://youtu.be/aqpxJ-5RIk8>

**GitHubvideo**

<https://www.youtube.com/playlist?list=PL5dllYcnnWYe9L9AIp4P6vQh8s8ptphis>

**LinkedInvideo**

<https://www.youtube.com/playlist?list=PL5dllYcnnWYc6D4_fRoslq6ddcsFauZvj>

**Hackerank video**    <https://youtu.be/sdDx89M_ww0>

**ResumebuildingRecording**

<https://drive.google.com/file/d/12XBkEwHD7o7sdHoeYIjUNzVt8vY0K_W7/view?usp=sharing>

**Resumebuildingvideo**

<https://www.youtube.com/playlist?list=PL5dllYcnnWYeJnmlD1Nrkqe-VeyJ8dHid>

**Annaconda/ Python Installation**    <https://youtu.be/tnBq7HXMSBE>

**LinkedInRecording**   <https://drive.google.com/drive/folders/1vFY8mXTdfhycse0MsZ7SriaO6jKstIjP?usp=sharing>

**LinkedIntemplate**    <https://docs.google.com/document/d/1Erf8TKpLOKxq4oHTLRUG1JtHKlJWq9GRq7zHbjpS2ZU/edit>

**postassessmentlink**

<https://docs.google.com/spreadsheets/d/1HR_bf4d4zgDnvfUrO8jJb2WQqvFRA9YqQj-QRbFgHU8/edit?usp=sharing>

**Lesson 1 -07th &08 August. 2021**

* OVERVIEW;

Understanding a problem / expectation of a client is important to any project.

Raising questions on data, is a required skill.

The new technologies available in the market.

Datascience entails different job profiles in the market.

* **How companies are utilizing these new technologies , in real time, to take them to the next level.**
* Machine learning , deeplearning, natural language processing, robotic process automation (RPA) and artificial intelligence nothing but a technology that can be applied in any industry or domain, which is why career transition into datascience has been easy.

By applying,

**Analytics** : the purpose of analytics is pulling out meaningful insights from data.

**Automation:** the purpose of automation is for revenue and time savings and to enable quick action and decision making

In 1950 Alok a businessman starts his business, he would sell different items . In 1965 , his sales dipped and he needed to effect some change . He approaches his mathematician friend Siddhu to help him with his financial issues.

Data is the raw material for any solution to be arrived at. Data is the new oil to provide solutions. Raising right questions on data is the essence of datascience.

For eg : Alok would have Volume of data with following parameters

**Date, Name , Product, Qty Rate Discount , Net price. Season**

This is the base theorem of probability and picked up to make a statement about future or early prediction or in other words, probability or early prediction

**Naïve’s base theorem** of probability could be one such statistical equation used to arrive at probable results.

P(a/b)= p(b/a) \* p(a)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

P(a)

If a customer is buying a blanket when the season is summer/w inter/ at this discount/ from this location/ at this price/ . By raising multiple questions on data, and understanding the domain/ impact of the data. A probable solution framework can be arrived at by predictive analysis and applying mathematical knowledge, leads to analytics. At this stage there is no application of automation.

Analysis here is only pulling out of meaningful insights from data and applying mathematical equations to arrive at solutions

**Naïve’s base theorem** of probability could be one such statistical equation used to arrive at probable results.

P(a/b)= p(b/a) \* p(a)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

P(a)

ML, AI, DL & NLP technologies are a subset of datascience. Lot of languages like python, r, sas etc. Python is simple yet has it’s own complexities. PYTHON & R are prime languages in datascience and AI

But this cannot satisfy all requirements. Issues of volume,scalability,revenue savings,generation of business,speed of delivery necessitates automation

**Types of analytics** : Fraud , Supply chain, Risk, HR, Sales,

**Learning python language is simple despite its complexities.**

**If goal is automation and analysis, one needs to get into python and for pure analysis, R would be more aligned.**

Lot of analysis can be done with python too.

Two prerequisites, best efforts and time.

**Three pillars**

Domain knowledge,Statistical Data Knowledge, Programming.

Datascientist combines all of the above and render great value to the client. Datascientist solves the unsolved problems,

**Learning grid** comprises of three elements : Learning rate, batches and iterations/ epoc

Iterations/ Epoch

Batches

Learning rate

The pace at which one learns. Human brain grasps in batches. Repetitions help in assimilation of information.

|  |  |  |  |
| --- | --- | --- | --- |
| **Designation** | **Purpose** | **Skills** |  |
| Data Analysis | Insights from the data to take business decision and actions | Excel, VBA, Macros, SQL, Basics Python, BI Tools - Tableau/PowerBI, Data Viz skills, Story Telling Skills |  |
| Data Analytics | Insights from the data to take business decision and actions | Excel, VBA, Macros, SQL,  BI Tools - Tableau/PowerBI, Data Viz skills, Story Telling Skills |  |
| Business Analyst | Insights from the data to take business decision and actions | Excel, VBA, Macros, SQL,  BI Tools - Tableau/PowerBI, Data Viz skills, Story Telling Skills | **DATA ANALYSIS** |
| SQL Engineer | Insights from the data to take business decision and actions | Excel, VBA, Macros, **SQL,**  BI Tools - Tableau/PowerBI, Data Viz skills, Story Telling Skills |  |
| ETL Engineer (Extract, Load, Transform) | **Data Collection +  I**nsights from the data to take business decision and actions | Excel, VBA, Macros, **SQL, NoSQL, Postgre, MangoDB,** BI Tools - Tableau/PowerBI, Data Viz skills, Story Telling Skills |  |
|  |  |  |  |
| Data Scientist | Insights + Predictions + Follow Up Action | **Python, ML, DL, NLP,**  Excel, SQL, Story Telling Skills, Tableau, Hadoop, Django, Flask |  |
| ML Enginner | Insights + Predictions + Follow Up Action | Python, ML, DL, NLP | **DATA SCIENTIST** |
| AI Enginner | Insights + Predictions + Follow Up Action | Python, ML, DL, NLP |  |
| DL Engineer | Insights + Predictions + Follow Up Action | Python, ML, DL, NLP |  |
| NLP Engineer | Insights + Predictions + Follow Up Action | Python, ML, DL, NLP |  |
|  |  |  |  |
| Data Engineering | to build **pipeline** to collect and store the data fro the cloud | Spark, Hive, Pig, Hadoop, MapReduce |  |
| Big data Engineer | to build **pipeline** to collect and store the data fro the cloud | Spark, Hive, Pig, Hadoop, MapReduce | **DATA ENGINEER** |
| Hadoop Enginner | to build **pipeline** to collect and store the data fro the cloud | Spark, Hive, Pig, Hadoop, MapReduce |  |
| Data warehousing Engineer | to build **pipeline** to collect and store the data fro the cloud | Spark, Hive, Pig, Hadoop, MapReduce |  |
| Data Architect | to build **pipeline** to collect and store the data fro the cloud | Spark, Hive, Pig, Hadoop, MapReduce |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
| Full Stack Developer (Data Science) | Inisghts + Prediction + Follow Action + to build pipeline to collect and store the data fro the cloud | Python, ML, DL, NLP,  Excel, SQL, Story Telling Skills, Tableau, Hadoop, Django, Flask, Spark, Hive, Pig, Hadoop, MapReduce | **DATA SCIENTIST & DATA ENGINEER** |
|  |  |  |  |
| Backend | To build Web Application | Django, and other WebApplication Frameworks. | **PYTHON DEVELOPER** |
| Python developer | To build Web Application | Django, and other WebApplication Frameworks. |  |
|  |  |  |  |
|  |  |  |  |
| DevOps | Inisghts + Prediction + Follow Action + to build pipeline to collect and store the data fro the cloud | Python, ML, DL, NLP,  Excel, SQL, Story Telling Skills, Tableau, Hadoop, Django, Flask, Spark, Hive, Pig, Hadoop, MapReduce, **Cloud Platform Knowledge - AWS, Azure, GCP,** | **DEV OPS ENGINEER** |
| MLOPS Enginner | Inisghts + Prediction + Follow Action | **Python, ML, DL, NLP,**  Excel, SQL, Story Telling Skills, Tableau, Hadoop, Django, Flask, Spark, Hive, Pig, Hadoop, MapReduce, **Cloud Platform Knowledge - AWS, Azure, GCP,** |  |

PYTHON DOWNLOAD

Python is Trending, easy, vastly used

PYTHON.ORG . Run in Windows 3.9 version 3.9.5

Download anaconda individual edition. Helps working locally offline

Run executable installer, just me option and choose destination folder.

Cmd python

Tutorial point.com / how to install python

**ANACONDA NAVIGATOR : INDIVIDUAL EDITION**, provide integrated development enterprise ( IDE)

**Learning log**

**01st lesson**,

**1)How companies are utilizing these new technologies , in real time**

* **Analytics** : the purpose of analytics is pulling out meaningful insights from data.
* **Automation:** the purpose of automation is for revenue and time savings and to enable quick action and decision making

2)Data is the raw material for any solution to be arrived at. Data is the new oil to provide solutions. Raising right questions on data is the essence of datascience

3)ML, AI, DL & NLP technologies are a subset of datascience. Lot of languages like python, r, sas etc. Python is simple and user friendly. yet has it’s own complexities. PYTHON & R are prime languages in datascience and AI

4)A datascientist one who solves unsolved problems.

**5)Different types of analytics**

**6)Different data related profiles or designations and their scope**

7)Three pillars of data science areDomain knowledge,Statistical Data Knowledge, Programming.

8)Learning grid comprises of three elements : Learning rate, batches and iterations/ epoch.

9)Google colab can be used online and not downloadable. Anaconda collection of different environment and provides various IDE. How to download Python.We shall be working on Jupyter notebook environment.

10) What is a codeshell. Writing basic code. Libraries include predefined functions or modules.

**Lesson 2- 14th August**

* As a datascientist one needs to focus on questioning skills. Lifecycle of a data scientist
* What is difference between datascience, Machine Learning,Deep learning,Natural language processing,Artificial intelligence.
* How to work on python

**Differences between AI, Machine Learning,Deep learning,Natural language processing**

**AI:** Feeding human intelligence into a computer For eg: Calculator, smartwatch. Understanding patterns and giving a response. Web applications gives in details for eg: eligibility for a loan. To execute AI we need machine learning.

<https://www.carwale.com/used/carvaluation/>

Against details input we get the output or valuation based on a pattern defined. Humans have given inteliigence to a machine and that is why it is called artificial intelligence. AI is not standalone It is proactive in nature and takes help of machine learning.

**MACHINE LEARNING** : Is a subset of AI which deals with learning patterns based on historical data. It uses algo

For eg: Spotify or Amazon.

It learns patterns by using machine learning by way of algorithm ( way of learning the pattern with the help of mathematical equations) AI takes help of machine learning through a structured format i.e data in a tabular format i.e rows and columns. For eg csv, xls, xlsx. Data needs to be supplied in a structured format.

DEEP LEARNING ; Subset of AI , learns patterns from structured and unstructured data, for eg: image classification or video applications or expressions . When data is unstructured format there is deep learning algorithms used.

**NLP;** where data is in a text form. Command given to machine and language processed. Processes language for prediction. For Eg prompts in predictions in mail/ smses etc. Machine makes predictions based on NLP . Text is also unstructured format. For eg; Alexa. Google Assistant Deals with pattern from text data, word prediction, sentence predictions,sentence analysis, grammar correction,machine translation, audio generation

**Komprehend.io** analyses emotions based on text through NLP operations.

Data structure is the main differentiator.

All technologies are based on data. Datascience is an intersection of analysis, prediction and automation

**Project** ; Has a startime and endtime / deadline. To complete or deliver the project we have several methodologies.

**Waterfall** : Speaks about delivering project in one shot. Disadvantage of lack of feedback and communication.

**Agile**: Delivering project in a phased manner in an incremental way or in sprint format. The advantage is the feedback option available. Various stages like design, raw materials, approvals. Trust factor is established in this mode.

**Lean six sigma.**

**SDLC: Software Design Life cycle**

* Planning, analysis, design, Implementation, Testing , Integration, Deployment, Maintenance
* Discussions on raw-material, source, resource, user acceptance testing.
* Irrespective of domain,the career transition to datascience is possible because of agile methodology.
* Imperative to create a trust factor with the client

**Life cycle of a data scientist**

**Agile Crisp: ( Cross Industrial Standard Process)**

* **Business understanding ( Goal),NLP, DL, AI.**

needs of the business or goal of a project to enhance revenue. Quicker processing of loan eligibility. Solving pain points of a client and impacting revenue of a client. This communication is crucial

* **Data Collection (Client) Statistics ,Domain knowledge**

Defining paramaters , age, name of customer, no of dependants, coapplicant details. Income details .

* **Data understanding Statistics ,Domain knowledge, Data visualization**

Approach data after understanding

* **Data preparation (Statistics ,Domain knowledge, Programming)**

Machine learns in a linear fashion with decoded value in a structured format lke in a row.

Data preparation is cleaning data and transforming data, transforming text to numbers. Machine will learn if definite patterns are provided

|  |  |  |  |
| --- | --- | --- | --- |
| **Age** | **Gender** | **Salary** | **Eligiblility** |
| **25** | **Female** | **10000** | **E** |
| **50** | **Male** |  | **NE** |
| **20** | **Female** | **5000** | **E** |

* **Model building ( Programming skills and statistics)** giving an input data and output data (segregating data 80 percent for training and testing)and correlation
* **Model training ( Math/ Algo, Statistics, programming)** providing sample question and answers to the machine
* **Model testing**: **( Programming)**  only question and answers are provided by machine.
* **Model evaluation ( Programming**) **( in sync with client expectation)** Output vs expected answers against solution sheet with accuracy of atleast 90 percent.If accuracy is matching with client expectation then it is passed
* **Model deployment** ( Programming, Cloud services)Upon matching with client expectation it is deployed to a website or local machine
* **Maintenace** ( Cloud Platforms Amazon Web service, Google Cloudservices Platform , Azure)

**Skills required**

Domain Knowledge

Programming Python – Analysis and automation or R analysis

Statistics,

Math algorithm.

Data visualization

NLP (data in text format)

DEEP LEARNING ( structured/ image/video)

MACHINE LEARNING (structured format)

AI (Feeding human intelligence not standalone)

**DATA SCIENCE**

Business Understanding

Model building

Model training

Model evaluation

Model maintenance

Model tesiting

Model deployment

Data Collection

Data Understanding

Data Preparation

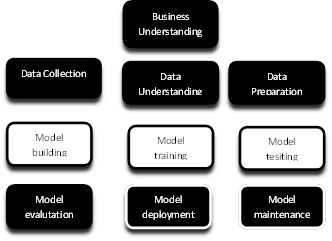
Life cycle of a data scientist

**SME SESSION:**

* Basic python codes.What is a code cell.
* If you have two statements in a code cell,without print command output of only last line is displayed
* Print without apostrophe will be treated as a variable and not as a text to be printed.
* Python is easier to manipulate, analyse and visualize data and is a language in trend.
* Def instead of function, instead of curly braces colon is used. In python no return type provided, automatically identified by compiler. Python is case sensitive.
* How to print command in same line or nextline.
* How to insert headings and titles using markdown/ heading.
* Commenting vide hash, is made ina code shell.
* How to rename a jupyter folder or filename

**Learning log – Lecture 2. 14th August.**

* Features and differences between Datascience, AI, Machine learning, deep learning , Natural Language Processing
* Different methodologies used for projects, eg. Waterfall, agile, leansix sigma. Agile crisp **( Cross Industrial Standard Process)**
* Lifecycle of a datascientist and various components in the same.
* Business Understanding, Data collection,understanding , preparation, Model building, training evaluation, deployment and maintenance.



* Skillsets required : Domain Knowledge,Programming Python – Analysis and automation or R analysis,Statistics, Math algorithm,Data visualization
* Basic python codes.What is a code cell.
* If you have two statements in a code cell,without print command output of only last line is displayed
* Print without apostrophe will be treated as a variable and not as a text to be printed.
* Python is easier to manipulate, analyse and visualize data and is a language in trend.
* Def instead of function, instead of curly braces colon is used. In python no return type provided, automatically identified by compiler. Python is case sensitive.
* How to print command in same line or nextline.
* How to insert headings and titles using markdown/ heading.
* Commenting vide hash, is made ina code shell.
* How to rename a jupyter folder or filename

**Lesson 3- 21st & 22nd August**

Types of Analytics

Data analytics

**Descriptive Analytics** Every project begins with descriptive analytics . It is pulling out meaningful insights from data.For eg Analysis of sales data with respect to parameters like time, region, etc. The tools basically needed are excel sheet, tableau, Power BI and python skills. Python is also used for data extraction. It is also called as diagnostic analysis, in order to answer a question of why, by asking multiple questions. This is done by the human basically the data analyst, by grabbing a definite pattern. Decision of execution of the analysis, too is taken by the human. Action too is taken by the human. It is a human centric activity.

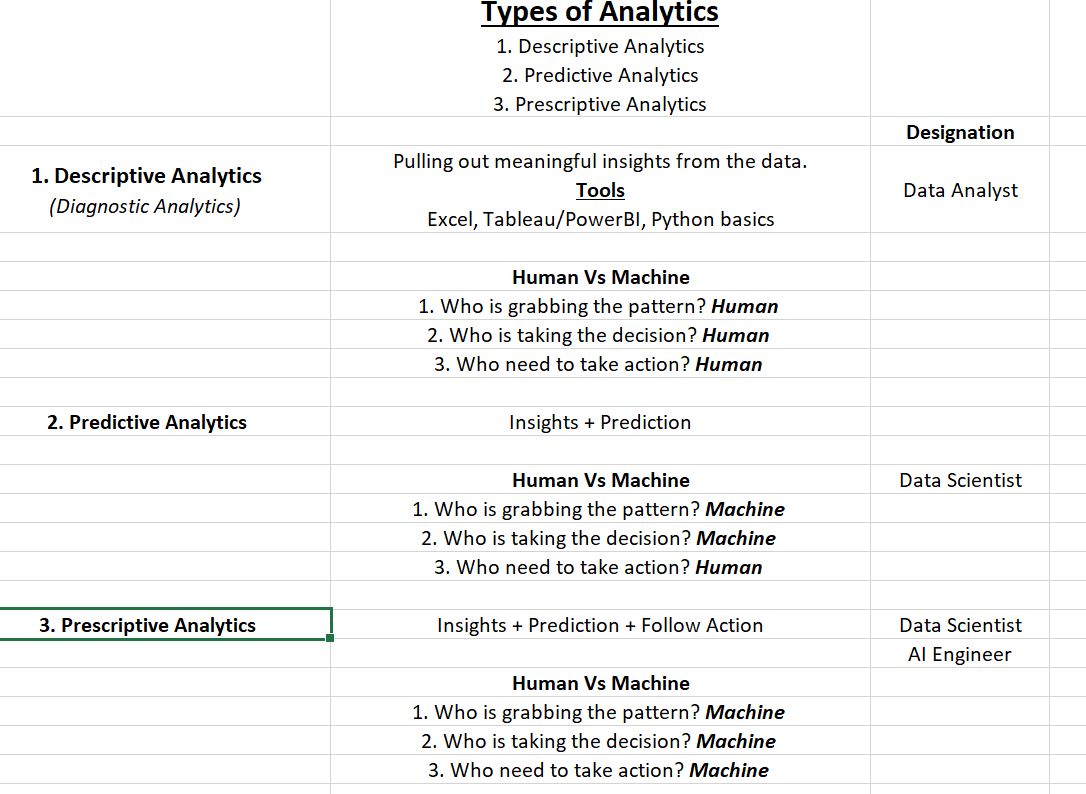
**Predictive Analytics**  Involves insights and prediction.

Based on inputs and patterns defined, machine gives the prediction

For eg; Carwale .com. Based on input and a certain pattern, a decision or prediction is determined by the machine and accordingly the result is provided by the machine. The action is taken by the human.Churn prediction is established by machine, based on pattern intelligence provided by the human.

**Prescriptive Analysis** Involves insights and prediction and also suggest followup action or suggestions based on the same. Data scientist and AI engineers are related profiles. It also involves Reinforcement layers based on analysis and learning from mistakes.

For eg: Self driving cars.



**Why Python**

* Python is easy to learn and read and code maintenance and intereactive.
* Python is case-sensitive.
* Python does not need a terminator
* Requires proper indentation.
* For path specification python users forward slashes
* One can use single double or triple quotes to represent string literals.
* Comments in programs by using # at the start.
* For analysis and automation. R language for analysis.

**Basics of python**

**Python data structures**: Like English grammar is structure of English language. List, tuple, sets and dictionary.

**Python libraries** numpy, panda , sklearn, tensorflow, matplotlib,seaborn,plotting of bars,piechart, and data visualization

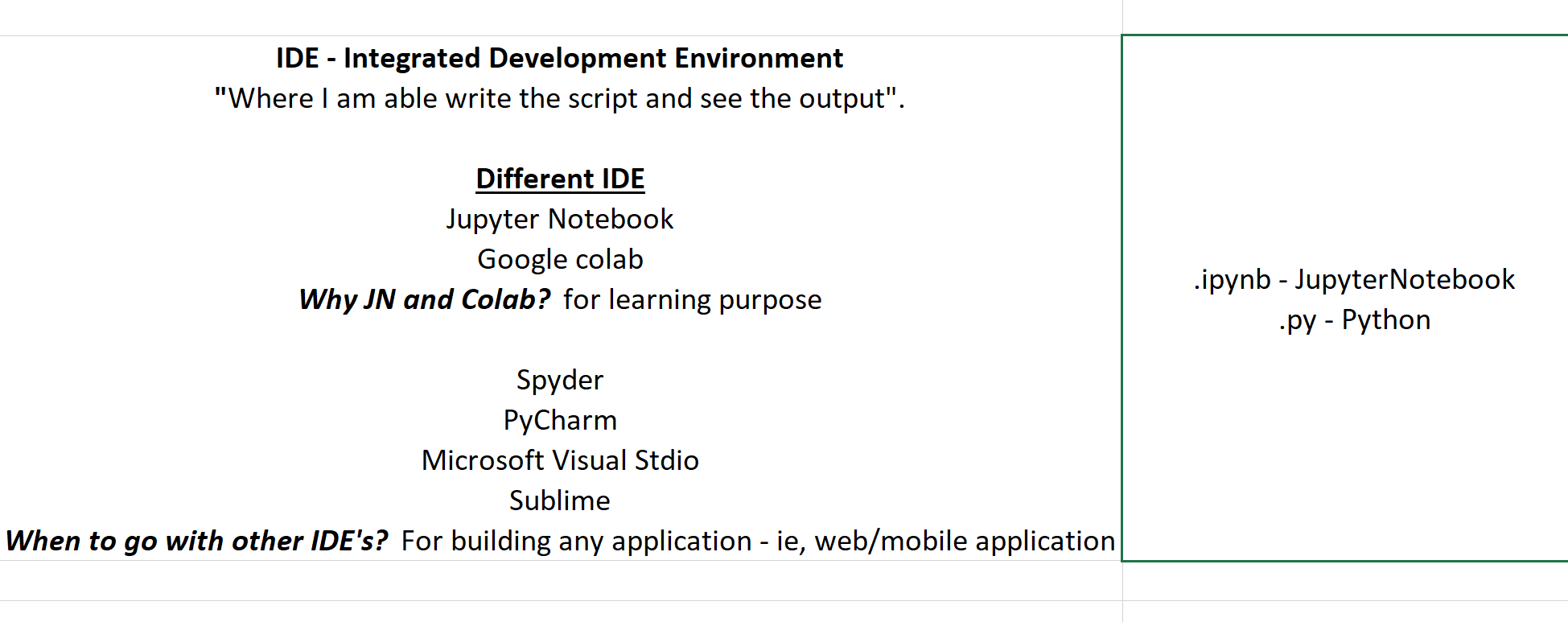
**Statistics**

**Basics of python**. Used for data analytics and automation . R is used for data analytics. Python enables analysis. Enables collaboration with datascientist team.

* Hello world
* String
* Number in python
* Functions and loops
* Conditional logic
* Data structures

We have different **Integrated Development Environment ( IDE**), where I am able to write script and see the output. Kernel is responsible for running the script in the background

Eg of IDEs Jupyter, Spyder, Orange. Google Colab, Pycharm, Microsoft Visual Studio Sublim



* In JN & Google Colab , one can learn from mistakes in coding immediately. GC provides for deep learning applications like video processing etc which calls for enhanced GPU. Google Runtime change allows for execution on google space . Jupyter runs on local machine, where processing time maybe slow.
* In spyder, script, file location window, visualization and output are available. In spyder there is no single cell for writing codes.
* Pycharm, similar to spyder id,you have different console views
* For application building we need a different format i.e .py. Default file extension for Jupyter is ipynb. For others it is .py . You can download a .py file and then can be uploaded on Google Colab. Jupyter will only run it in ipynb file format. Py enables manual reading
* JN and google enable better learning

**Anaconda navigator** is a collection of different IDEs.

**MARKDOWN METHOD** for creating titles.

**Shortcuts**

**Short cut method** X to cut,Z to revert,B, Create new cell below current cell, A for cell above current cell. Typing H gives list of all short cuts.

When we click outside cell it is blue , inside cell color is green

Type outside 1 # for heading , 2 for subtitle

**List of shortcuts**

F: find and replace

Ctrl-Shift-F: open the command palette

Ctrl-Shift-P: open the command palette

Enter: enter edit mode

P: open the command palette

Shift-Enter: run cell, select below

Ctrl-Enter: run selected cells

Alt-Enter: run cell and insert below

Y: change cell to code

M: change cell to markdown

R: change cell to raw

1: change cell to heading 1

2: change cell to heading 2

3: change cell to heading 3

4: change cell to heading 4

5: change cell to heading 5

6: change cell to heading 6

K: select cell above

Up: select cell above

Down: select cell below

J: select cell below

Shift-K: extend selected cells above

Shift-Up: extend selected cells above

Shift-Down: extend selected cells below

Shift-J: extend selected cells below

Ctrl-A: select all cells

A: insert cell above

B: insert cell below

X: cut selected cells

C: copy selected cells

Shift-V: paste cells above

V: paste cells below

Z: undo cell deletion

D,D: delete selected cells

Shift-M: merge selected cells, or current cell with cell below if only one cell is selected

Ctrl-S: Save and Checkpoint

S: Save and Checkpoint

L: toggle line numbers

O: toggle output of selected cells

Shift-O: toggle output scrolling of selected cells

H: show keyboard shortcuts

I,I: interrupt the kernel

0,0: restart the kernel (with dialog)

Esc: close the pager

Q: close the pager

Shift-L: toggles line numbers in all cells, and persist the setting

Shift-Space: scroll notebook up

Space: scroll notebook down

Edit Mode (press Enter to enable)

Tab: code completion or indent

Shift-Tab: tooltip

Ctrl-]: indent

Ctrl-[: dedent

Ctrl-A: select all

Ctrl-Z: undo

Ctrl-/: comment

Ctrl-D: delete whole line

Ctrl-U: undo selection

Insert: toggle overwrite flag

Ctrl-Home: go to cell start

Ctrl-Up: go to cell start

Ctrl-End: go to cell end

Ctrl-Down: go to cell end

Ctrl-Left: go one word left

Ctrl-Right: go one word right

Ctrl-Backspace: delete word before

Ctrl-Delete: delete word after

Ctrl-Y: redo

Alt-U: redo selection

Ctrl-M: enter command mode

Ctrl-Shift-F: open the command palette

Ctrl-Shift-P: open the command palette

Esc: enter command mode

Shift-Enter: run cell, select below

Ctrl-Enter: run selected cells

Alt-Enter: run cell and insert below

Ctrl-Shift-Minus: split cell at cursor(s)

Ctrl-S: Save and Checkpoint

Down: move cursor down

Up: move cursor up

**SME session**

**Difference between print and input function :print to give output as it is &input statement takes input and gives output simultaneously** print('Welcome')

firstname=input('Enter your first name')

middlename=input('Enter your middle name')

lastname=input('Enter your last name')

print(firstname,end="\_")

print(middlename,end="\_")

print(lastname)

Variable does not require giving quotes inside brackets and print command requires strings to be given in quotes. Variables are like containers to store values of different types

**Variable naming convention**

* Variables names cannot begin with numerals
* Variable name cannot have spaces but underscore can be used, before or between variable name.
* Predefined keywords cannot be used as variable name. for eg . Function name like print.
* In python, names are case sensitive.

Predefined words are encased in brackets.

In python by default datatype is string. To typecast i.e to convert one datatype to anotherPredefined data types it has to be typecast.

Integer = int

Decimal=float

String=str

**Data structures**:Tuple, list, sets, dictionary, array.

Learning log .

Session 3 -21st August, 2021

1. Different types of Analytic – Descriptive (Data Analyst), Predictive ( Data Scientist), Prescriptive ( Data Scientist and AI Engineer. How scope of each analysis differs / progresses.
2. Why Python? Features of python.
3. Anaconda provides a collection of IDEs that is where I script and get my output.
4. Jupyter Notebook and Google Colab are used for learning purposes and their various features. File extensions being ipynb and py respectively. Ipynb is like an encrypted form and needs the corresponding application to read whereas .py files can be easily read. JN and google enable better learning
5. How each is different from the other and how for advanced mathematical calculations higher CPU processing power is required. How other IDEs are used for building web or mobile applications.
6. Exploring shortcuts in JN and GC
7. **Difference between print and input predefined function.**
8. **How to print in a single line or with user specified characters**
9. **What is a variable- A container for inputs**
10. **Variable naming conventions**
11. **Datatypes, string (default),integer,float(decimals),boolean, NoneType.**
12. **Typecasting of datatypes. Datatype by default for all inputs is string and can be changed.**
13. **Datastructures- List,Tuple, Sets,Dictionary**

Learning log

Lesson 4; 28th August

**Variable:**

* Based on the content of print statement,single,double or triple quotes are used.
* Three important datatypes in python
  + Int 32 (no decimal values)
  + Float 32.23 (with decimal values)
  + String enclosed within quotes can be single or double quotes or triple quotes.
* String datatype cannot be stored without quotes
* Variable is like a container can store multiple values and datatypes.
* A list datastructure can contain heterogenous datatypes.
* After a variable is created, it lends itself to manipulation.
* Keyword len returns length or number of values in a variable.
* Slicing helps accessing a portion from the variable i.e filtration.
* Slicing is done with the help of index number within square brackets.
* In PYTHON index number begins with zero.
* In indexing, start position is inclusive and end position is exclusive
* Variable needs to be store data in form of a structure/datatypes
* Variable naming conventions does not allow spaces and function keywords are best avoided.

**Strings and methods/functions**

* Difference between datatypes string, int and float.
* What is a string?
  + Any values within quotes is treated as a string
  + Concatenation,indexing and slicing
* Comma or + is used for concatenation and works for string datatype only
* Escape sequences to be included within quotes /n- nextline /t- tabspaces
* Manipulate string with methods
* Methods/Functions used to operate certain operation on a variable & are enclosed in brackets.
* When I want a perform operation on a variable is a function.
* Three types of functions
  + Predefined functions gets installed when we install an IDE. and is a part of the IDE
  + User defined functions customised functions where user writes his own set of instructions to be performed
  + Anonymous or Nameless functions used for limited operations
* It is not recommended to use keywords as variable or user defined functions.
* Right hand side gets stored into left hand side, = is an assignment operator, RHS assigned to LHS.
* No limitation to number of values that can be given in a print format function
* Find a string within a string with the find() predefined function, output as starting index number.
* In an index first postion is stored at 0 ,start index inclusive, end index is exclusive.
* Format of an Index[start,end,skip]
* Negative index start is -1 instead of zero
* Difference between upper and isupper functions

**Learning log : Lecture 5 : 29th August.**

1. Functions are used for resusability of code
2. Def is keyword for definition of a user defined function
3. Abs function for Conversion of negative values to positive using abs function
4. Chain of methods: X=abs(int(input=’Enter number’))/ print(abs(int(input(‘Enter a negative value’)))
5. Functions and loops, writing own functions When we create a function, the variables therein are called arguments. But when the function is called the same are called parameters.
6. Return statement completes a statement. A function can have multiple print statement but only single return statement.
7. A return statement terminates the code. So print statement should precede return statement.
8. When we create a function arguments are passed, when we call the function the values are called parameters
9. Positional and default arguments.
10. **Variable length arguments** can be provided by putting an asterisk before the last argument. We cant have two variable length arguments.
11. Variable keyword argument.
12. A variable keyword is defined by two asterisk. When it is a variable keyword, the keywordname passed in parameter should be same as argument name with an extension.
13. Variable length and keyword arguments. Only one of each in a function with length and keyword arguments. This order is this: first comes Standard arguments, then comes \*arguments (known as \*args) and then comes \*\*arguments (known as \*\*kwargs)
14. What is a stepsize in for loop? Incremental value for running of loop,default being 1
15. Control flow of statement, looping conditions
16. Conditional logic- compare values
17. While true condition
18. Create a user defined function with for loop and if else
19. Breaking out of a loop.
20. Function can have multiple print statements but only one return statement..Without return statement also, function will work and can be achieved with print statement too.Print allows multiple statements and once return statement is provided code gets terminated. So print statement should precede return statement should be given in the end only. If there are multiple return statements only the first statement will be executed.

Extra notes for Lecture 5 -29th August.

* **user defined function** has a name and when user types the function name the return values are displayed

**def get\_details(): \*\* defining a function**

**return” Some values “ \*\*terminates code**

**get\_details() \*\* calling of a function**

3) When we define a function **arguments** are passed

def add\_numbers(num1+num2+num3)

return "The result of formulas is",num1+num2+num3

add\_numbers()

The parameters get stored in corresponding argument in same position.

**Positional arguments**

**Defined positions for arguments in a user defined function. A certain specified positions of parameters and arguments are followed**

**get usr\_dtls(idno,name, age, Company):**

**While calling the function we have to provide corresponding parameters. Any argument which is common can be defined in the function creation itself**

**def get\_usr\_dtls(id\_no,name,age,Company='Google'):# default argument when function creation**

**print('ID:',id\_no, 'Name:',name,'Age:',age,'Works at:',Company)**

get\_usr\_dtls(1, ‘Suresh’,25,’Google’)

get\_usr\_dtls(1, ‘Jignesh’,28,’Google’)

get\_usr\_dtls(1, ‘Sudha’,34,’Google’)

get\_usr\_dtls(1, ‘Pratiksha’,22,’IBM’)

**user defined functions are used to repeat same function with change of values only**.

* **Variable length arguments** can be provided by putting an asterisk before the argument. We cant have two variable length arguments.

**get usr\_dtls(idno,name, age, \*Company):**

**Here multiple values can be passed for Company**

* **Variable keyword argument.**

**A variable keyword is defined by two asterisk. When it is a variable keyword, the keywordname passed in parameter should be same as argument name with an extension.**

**There is a order of writing arguments when you are using \*arguments and \*arguments. This order is this: first comes Standard arguments, then comes \*arguments (known as \*args) and then comes \*\*arguments (known as \*\*kwargs**)

**which is why the following will not work**

**def get\_usr\_dtls(id\_no,\*name, age,\*\*Company):**

**def get\_usr\_dtls(id\_no,\*name, \*\*age,Company):**

**whereas this will work.**

**def get\_usr\_dtls1(id\_no, Company, \*name, \*\*age):  
  print('ID:',id\_no, 'Name:',name,'Age:',age,'Works at:',Company)**

**get\_usr\_dtls1(3, 'Google', 'Shaila', 'Angela', 'Sid', age\_1=28, age\_2=45, age\_3=62)**

**Run in circles**

**Loop condition** to increment by a certain value , within a range

For I in Range( startnumber ,end number):

**range(start, stop[, step]) -> range object, stepsize i.e how much it has to increment and pass into the i variable**.

For, if elif, else loops

Syntax of while loop

While condition , set of instructions, increment

**Learning log:**Lesson 6 : 04TH & 5th September

1. Why datastructures?
2. Because like any language has a grammar or structure, similarly python has it's structures.
3. 4 datastructures in python, list, set, tuple,dictionary
4. Propeties of List: It is indexable, sliceable,can contain duplicate values,one-dimensional,heterogenous datatypes ,mutable ie. values can be changed anytime in future. Empty list is Initialized by [].
5. Properties of Sets:Set is heterogenous, one dimensional, not indexable, not sliceable, no duplicates, not mutable, unordered.To initiate set -> set(), to store values in set set {},set is not subscriptable cannot be indexable or sliceable.All datatypes have their own methods->set.function,set cannot return duplicates,in order to get **unique values** set can be used.Pandas cannot add set datastructures into a dataframe.Sets are used in NLP related applications.
6. Properties of Tuple: Add value () Can contain duplicate values, heterogenous, one dimensional Is indexable, sliceable, immutable if you dont give brackets it will take default as tuple.Tuple to be used when we want to **store values that cannot be changed**
7. Properties of **Dictionary also called - Key:Value Pair :Two dimensional datastructure,**not indexable,Key:value Pair,Empty dict initiated by ={},cannot have duplicate keys since it will get overwritten, can have duplicate values
8. Pandas library. Lends to manipulation of dataset
9. Importing dataset into Jupyter notebook, by either uploading/ tracing from destination folder.
10. How to use separator in split method in a string.
11. Values in a dictionary can be a dictionary, it is called nested dictionary
12. Nested functions

**Lecture 7 -05th Sept, 2021**.

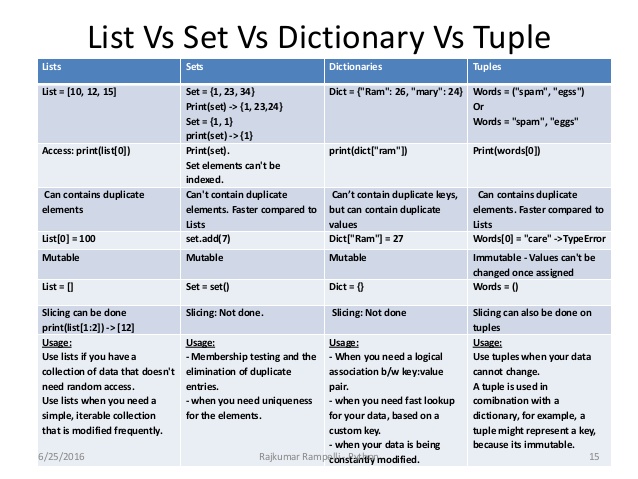
1. Data manipulation possible only if stored in variable.
2. index id not counted as column for data analysis
3. Attribute doesnt end up with open and end parenthesis, for eg: shape
4. Top n number of records,head(), bottom n number of records tail().Script pd.set\_option('max\_rows',20) # max rows None, would display all values or you could limit display by using set\_options configurationscript
5. Rows= Observations /Records/Data point.
6. Columns= Features/Parameters
7. How to identify pandas datatypes ie

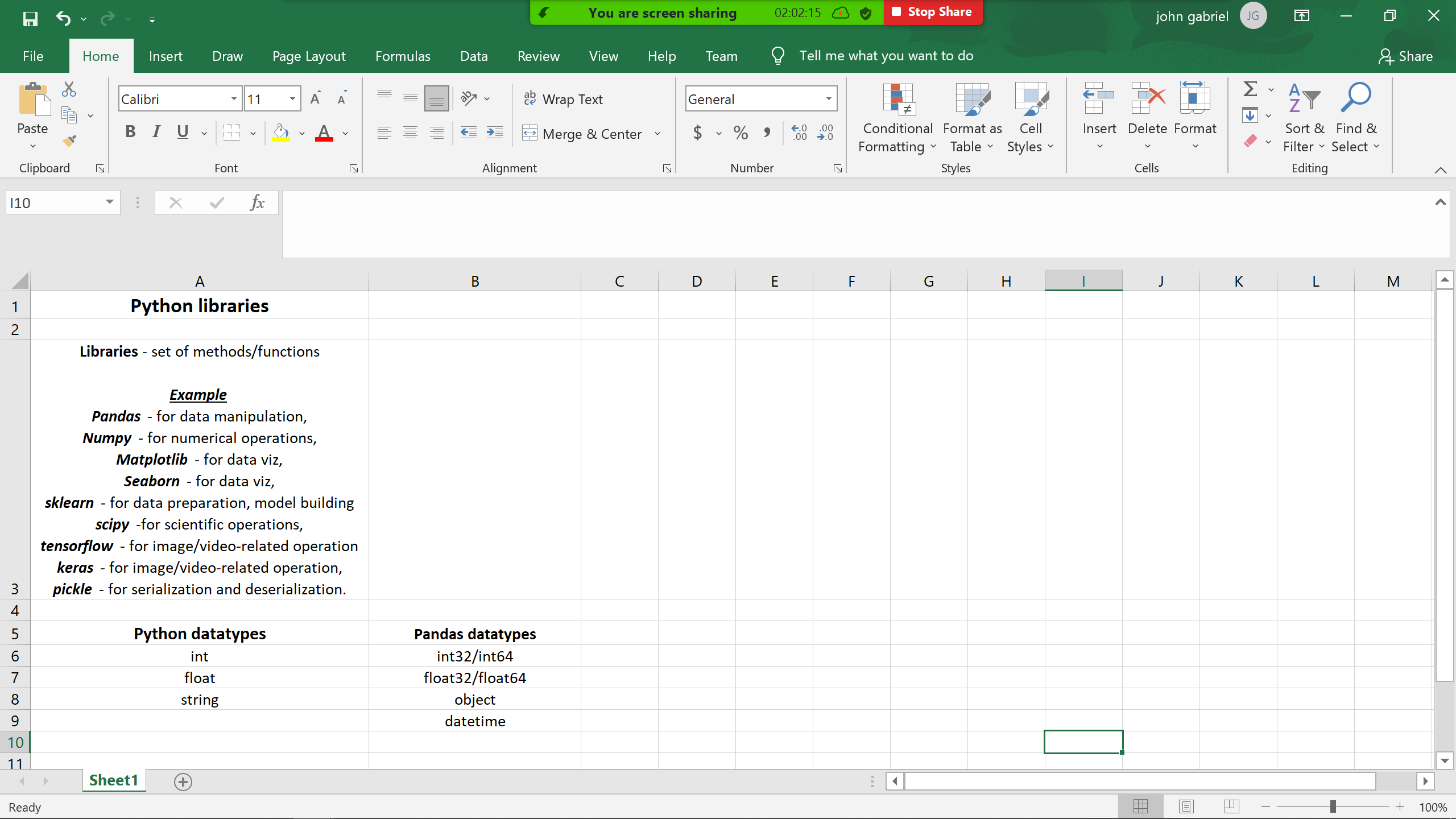
Int- Int32/Int64

Float- Float32/64

String=object. Any special character is treated as object in pandas.

1. How to check nullvalues . Percentage of missing values
2. When information is incomplete clustering or grouping is not possible, so null values are generally replaced by certain values.
3. variable[.]info his would return details of each column. This shows datatype, and memory usage data.
4. nunique()function.- sum of unique values.
5. Unique() function lists the unique items.
6. Grouping of data.
7. Value counts gives frequency
8. Statistical report can be got from pandas[.describe(include=’all’),describe gives report for numerical values int and float.
9. Group by first feature categorical and number based feature and mathematical operation





18th September 2021

Resume building session:

What interviewers are actually looking for

Avoid unnecessary details

Resume structure

**Websites for resume building**.

Novoresume

Resume.com

Google docs

Canva.com

Great sample resume

CV is more detailed than resume.

**Professional Title**

**Non technical skills** :Communicaion, leadership, project management critical thinking.

**Freshers Experience**

Internship

Non techncial jobs

Experience in building a particular project

**Don’t**

Make lengthy

Avoid gramattical mistake

Don’t lie in a resume

Never comment negatively about former company or employees.

Lesson 8 :18th September- Learning log

1. Converting one datatype to another is called typecasting.
2. If there are symbols like $ it does not allow direct typecasting.
3. purchase\_data['item\_price']=purchase\_data['item\_price'].str.replace("$","").astype('float')
4. Data cleaning is a very important exercise.
5. Replace function used with typecasting.
6. Import warning and ignore warnings.
7. We cannot delete two columns at once.
8. Argument inplace true impacts original data & default value is false.
9. Where inplace is not there create a new variable and assign values.
10. Drop method used to delete multiple labels based on rows or columns
11. Axis = 0 --> Rows || Axis = 1 --> Columns
12. How to select specific columns from a dataset.
13. How to sort data using sort\_values method.
14. No separate descending argument in sort\_values, ascending=false to be used.
15. Drop duplicate method.
16. For selecting two columns, two square brackets to be used.
17. Pandas is having 2 datastructures: Series - 1D DataStructure.DataFrame - 2D DataStructure
18. Group by can be applied first on categorical values and subsquently on numerical value, followed by the mathematical operation required.
19. When we want to drop a column we can use axis=1 , else if one wants to read values of a row in a column there axis =0
20. Imputation replacing null values with some values.
21. Apply is a predefined function,but not standalone takes support of predefined or anonymous functions.
22. Anonymous functions are defined using lambda keywords
23. A lambda function can have multiple arguments but only one expression
24. Syntax of lambda function is lambda argument: expression.
25. Filter function with lambda
26. Filtering and sorting with index locator iloc function.
27. Merge function.
28. mergedata=pd.merge([sales\_2017,sales\_2018,sales\_2019])
29. full\_data=pd.concat([sales\_2017,sales\_2018,sales\_2019])
30. How to Convert data to csv , excel sheet etc.

**Learning log 9 – 19th Sep,**

1. Iloc and loc are operators iloc is an index locator.
2. Syntax =iloc[4:10,1:3] iloc[rowstartindex:rowendindex, colstartindex:colendindex)
3. lloc for picking up wih column names instead of index, can be used where there are many columns,
4. Filtering can be done with negative indexing too
5. Negative index begins with-1 and positive index begins with 0.
6. Pivot table, cross tab, groupby serves same purpose.
7. Group by discrete/categorical data,
8. Pivot tables helps in presenting the aggregation process in a structured format,where there are more than one columns.
9. Cross tab function cross helps understand relationship between categorical data
10. In case of cross tab there is no data parameter so it has to be directly called in the code. In case of pivot table there is a reference to a dataset. It helps in building a frequency table, categorical columns to be picked up both in index and columns.
11. Margins in crosstab helps get total of values in columns and rows.

|  |  |  |
| --- | --- | --- |
| Group by | Pivot | Cross tab |
| Groups data in different categories |  | Pivot and cross tab similar to eachother and used for the same purpose |
| Data is grouped based on 1 or several variable and analysis is perfomed on individual groups |  | This method is used to compute a single cross tabulation of two or more factors |
| It is recommended to apply group by functions on categorical columns so that the function could be meaningful |  | Computes frequency table of the factors unless an |
| If you notice while grouping more than 1 column, onecolumn has more thatn 2 categories and theother column has more than 2categories itisalways better to go with pivot table because the format of table display it would better compared to groupby | This data is much easier to visualise when it is made into a pivot table. More presentable than groupby |  |
|  |  |  |

**Data visualisation.**

1. For easy understanding of data, it save more time because we can quickly analyse data, for presenting / showcasing /story telling and dataquality check.
2. There are various business intelligence tools , with python we rely on matplotlib, seaborn and plotly. These are libraries.
3. **Python based visualisation libraries**
   1. Matplotlib- more customisation on my graph. helps in customization of visuals and transferring of images on to power point. Help easier undersatanding and effective story telling. Pyplot is a sub package of matplotlib. %matplotlib inline

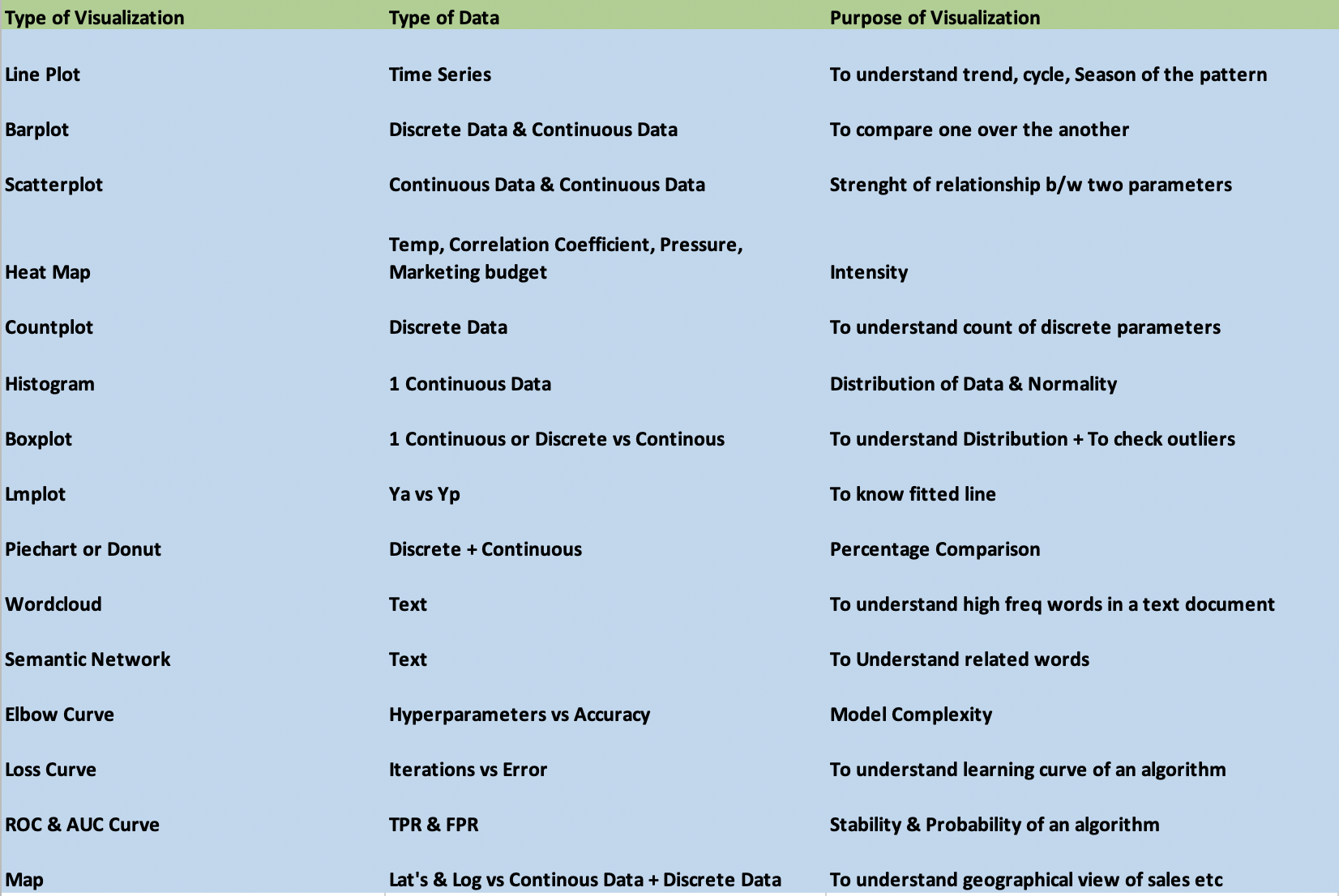
magic command to view charts within the jupyter page.

* 1. Seaborn- for beautification and exclusive graph types
  2. Plotly- to create interactive charts

1. Charts are chosen based on type of analysis needed.Univariate analysis, bivariate analysis, multivariate analysis.
2. Discrete and continous data , univariate and bivariate data and how that impacts the choice of graphs /charts. Discrete data not measurable, does not have units,just counts it cannot be segmented. Continuous data is measurable has units and can be segmented.
3. How to export charts to powerpoint
4. What are ticks in pyplot?
5. How seaborn is used to beautify charts built on the base of matplotlib.
6. Seaborn has advanced plots like countplot,violinplot,pairplot,catplot,relplot,heatplot
7. How to build subplots . Building multiple plots in a single plotting aarea
8. Sns Samuel Norman Sams , sns used as alias for seaborn.
9. Basis of choice of charts.

|  |  |
| --- | --- |
| **Discrete** | **Continuous** |
| Not measurable in units only count | Measurable in units ,can be in fractions |
| Cannot be segmented | Can be segmented |

|  |  |  |  |
| --- | --- | --- | --- |
| Variate | Type of chart | X axis | Yaxis |
| Univariate | Histogram | Continuos |  |
|  | Pie | Discrete data along with counts |  |
|  | Countplot | Discrete data (Categorical feature) |  |
| Bivariate | Bar | Discrete | Continous |
|  | Scatterplot | Continuos | Continuos |
|  | Boxplot (Five value chart) | Continuous |  |
| Multivariate |  |  |  |

****

**Brief summary of learnings till this stage**.

Life cycle of a data scientist, type of analysis, descriptive & predictive, basic of python, creating a variable , different functions, for loop and while loop and if statement, python datastructures, python libraries, collection of methods and datatypes. Pandas help data manipulation, shape of data,groupby, paranthesis, filtering, iloc, indexing, deleting, deleting multiple columns.

Summary: One needs to speak to data through programming

The lifecycle of a datascientist, Start with a project with Descriptive , predictive and prescriptive analysis

Basics of python programming, importance of creating a variable, userdefined functions, predefined functions and anonymous functions. For loop, while loop, if and ifelse statements,

Python datastructures, list, set,tuple dictionary. Dictionary can be made into a dataframe format i.e rows and columns.

Python libraries, collection of methods and its datatypes. Several type of libraries for different purposes

Pandas helps datamanipulation and analysis How to read and raise multiple questions on data, different methods, attributes, filtering and sorting grouping, replacing, indexing, delete data,basically data manipulation.

**Learninglog Lecture 10- 25 th Sept**

* List comprehension : Writing code in one line.
* Plotting tools in plotly
* **Univariate Analysis** 
  + Histogram to be applied for continuous data distribution
  + Pie chart shows percentage wise comparison of
    1. discrete data (Discrete Vs Continous data Discrete data - it is not measurable(no units, just counts)/it cannot be segmented. Continous data - it is measurable(it has units) / it can be segmented)
    2. returns value\_counts
  + Countplot from seaborn for beautification, for discrete datatype.
* **Bivariate analysis**
  + Barplot ( x axis =discrete data and y axis continuos)
  + Sns.barplot gives beautified barplot.
  + Seaborn give better accurate plot and display than matplotlib. Matplotlib for customization
  + If we don’t pass y axis in barplot, will not give an error but will not be presentable data.
* List comprises of elements of different datatypes. List comprehension is a shorter way of defining a list.
* Syntax for list comprehension [expression: for loop condition]
* Differences between Matplotlib & Seaborn
  + Both libraries are visualisation library.
  + Seaborn is based on matplotlib,interactive charts and advanced plots in seaborn. The important plots are
    1. matplotlib: lineplot, piechart, histogram,barplot,boxplot
    2. only seaborn:countplot,violinplot,pairplot,catplot,relplot,heatmap.sunburst
    3. subplot :Building multiple plots in a single plotting area
* plt.figure( width,height).
* Seaborn has Its own set of datasets which can be used. sns.get\_dataset\_names () for getting names of datasets in seaborn

anagrams,anscombe,attention,brain\_networks,car\_crashes,diamonds,dots,exercise,flights,fmri,gammas,geyser,iris,mpg,penguins,planets,taxis,tips,titanic

* How to plot subplots
  + fig,ax=plt.subplots(2(rows),4(columns),figsize=(9,9))

**Lesson 11-26th September.**

* Plotly.com
* Scatterplot: to understand linear relationship/ association between two continuous datatypes.Two axis needs to be continuous. Can be inverse or proportionate relationship that is can be negative or positively correlated. Scatterplot shows linear relation
* Correlation matrix., Correlation matrix speaks of strength of linear association between features. Dataset.corr.
* Heatmap ( speaks of intensity of correlation )
* Mulitvariate analysis : Linear association between multiple features.
* Summary of correlation between all parameters given by Pairplot in one shot.
* Boxplot plots continous data. Iti is a fivevalue plot. Used to understand data distribution and to detect outliers i.e variance from general data. The describe (include=’all’) helps understand the values better.
* How to plot subplots,defining indexes for subplots
* Impact of hue feature in countplot.
* In matplotlib, (axis, type of graph,column to be plotted)
* In seaborn( sns.typeof graph(column,dataset,axis)
* Boxplot visualisation the percentile points and outliers.
* Outliers are those points which behave differently from normal data.
* Ways to detect an outlier
  1. Boxplot (Visualisation)
  2. Zscore (Statistical)
  3. IQR -Inter quartile range (Statistical)
* What is percentile.
* IQR beween 25 and 75 percentile. Is inter quartile range
* Iqr method is a statistical concept, and boxplot is a visualisation concept to identify outliers
* Difference between mean and median and how outliers impact mean values. If data has outlier,true representation would be median instead of mean. Outlier are extreme elements in adata. 50th percentile is the median, average is mean. If mean =median, there are no outliers. Ways to detect outlier using IQR using numpy.percentile
  1. Sort dataset in ascending order,
  2. identify q3 and q1 percentile ,
  3. identify IQR
  4. find lb and ub
  5. any value s outside of lb and ub is an outlier
* iqr=q3-q1
* lb=q1-1.5\*iqr ( lower bound)
* ub=q3+1.5\*iqr ( upper bound)

When there is outlier median (middle value) , and when there is no outlier mean =median

Different plots Lineplot,piechart,histogram,barplot,boxplot, catplot,recplot, heatmap,countplot, violinplot, pairplot

**Learning log Lecture 12 03rd October**

1.Numpy library for performing numerical operations,array operations.

2.For creating dummy entries.

3. Size of list can be changed, but array size cannot be changed values cannot be added to an array/numpy. It can be incerased by creating a new array

5. In numpy array size is fixed, New values cannot be added, but some arithmetic operation and resize, can be done

6. List, we cannot perform addition/ or any operation to the values present inside.( elements)

7.Creation of multi dimensional array.

8.Converting integer to float using numpy.

9. Three dimensional data, three square brackets to be enclosed.

10. Converting list to array.

11. np.array can be converted to list but the dimensions cannot be assigned to the list

12. Array variable.size gives number of elements.

13 . Creating nan with numpy.

14. Conversion to Nan with numpy can be done for float values only.

15.Increase number of samples by repeat method

16.Concept of arange in numpy. (np.arange(1,31,dtype = 'float')

15. Argmax works for one dimensional array and returns index

16, In array. Reshape ,nature of dimension not changed

17,To change dimension flatten is used. It collapses to one dimension

18.Random number generation np.linspace(start = 1,stop = 10,num=50)

19.random\_num\_1 = np.random.rand(nrows,ncolumns) # 0 to 1

20.random\_num\_2=np.random.randn(10,10)#normally distrubuted numbers

21.np.random.randint for generating random numbers as integers

22.np.random.seed(123) helps in fixing the random numbers.

23.Creating arrays with specified rows and columns and specified element eg np.zero, np.ones

24. NORMAL/GAUSSIAN DISTRIBUTION: follows an empirical rule:

* 68% of the datapoints, will fall between -1SD to +1SD.
* 95% of the datapoints, will fall between -2SD to +2SD.
* 99.99% of the datapoints, will fall between -3SD to +3SD.
* Beyond these values is an outlier.

25. Normal/gaussian distribution graph follows a bell curve

26.Variance= how far a datapoint is spread from mean

27.Standard deviation= square root of variance.

28. mean= sum of all datapoints (xi)/ number of datapoint ( mew)

29. variance=sigma squared=(datapoint-mean) squared /value for all datapoints i.e net variance

30. variance=sigma squared=(distance) squared /value for all datapoints

31. Z score any data point beyon 3SD is an outlier.

32.Normal distribution points are better to analyse.

33.Types of distorted gaussian distribution. Skew is measure of assymetry. The presence of outliers lead to asymmetry.

* left skewed
* right skewed,

Normal : mean=median=mode

Positive/right skewed = mode, median mean

Negative/left skewed =mean, median, mode

34. Kurtosis depicts distortion in terms of height.

**Learninglog Lecture 13 -09th October 2021.**

1. Need for statistics:To make a decision in the present, based on historical data. Statistics is about studying collected data.
2. Descriptive & Inferential Statistics.
3. **Descriptive Statistics**
4. Sample should be = to population in lines of properties but can vary in size.
5. Sample is required because population size can be large.Sample collections requires domain knowledge.To understand data we look into EDA.
6. EDA= **EXPLORATORY DATA ANALYSIS** – Initial investigation of data
   * **Measures of Central Tendency** – Mean Median Mode

If I want to understand the central value of the data  
Mean - when there is No outilers.  
Median - in the presence of outliers.  
Mode - to look into the most occuring value/category

* + **Measures of Dispersion** –Variance, Std deviation, range
    1. (How well the values got deviated from the central value/mean)  
       Variance - units will be squared terms  
       Why Std.Deviation? - Inorder to normalize the variance.   
       Why Range? Max - Min. Not mostly suggested.

**Point to Remember**  
If the variance increases, the quality of the decision/prediction decreases.

* + **Normal Distribution**  
    ND follows the Emperical rule and bell curve. Symmetrical curve. For eg. Age and parental dependency / Mobile phones versus price.  
    For 1SD(-1SD to +1SD) - 68% of the data points had fallen between the range.  
    For 2SD(-2SD to +2SD) - 95% of the data points had fallen between this range.  
    For 3SD(-3SD to +3SD) - 99.99% of the datapoints had fallen between this range.
  + **StandardNormal Distribution.**
    1. Standard Normal Distribution where mean=0 and SD=1
    2. Both follow empirical rule with SND sample is restricted.
    3. When data is not normally distributed log.SND will try to get better values.
  + **Skewness & Kurtosis**.

Measures lack of symmetry in data distribution .If tail of distribution is towards right it is right skewed ( Median-> Mean-> Mode), if left left skewed ( Mean-> Median-> Mode).

If there is Skewness in my data - it impacts the ACCURACY of my parametric model.  
If there is KURTOSIS in my data - it impacts the STABILITY/PRECISION of my parametric model

Accepted skewness levels

* + 1. 0.5 to 0.5 normal;
    2. -1 and -0.5 or between 1 to 0.5 moderately skewed,
    3. less than-1 more than 1 highly skewed

**Kurtosis** speaks of tails of distribution High Kurtosis (lepto) mean heavy tails, low kurtosis means light tail (platy).

Mesokurtosis=3 Leptokurtosis > 3 Low or Platy < 3

**OUTLIERS  
"The most extreme and most least values"**

1. Variance are in squared units to normalise the same we use standard deviation.
2. Variance= square of sd = xr- central value i. e mu / number of value
3. Sd= square root of
4. Range is maximum- minumum value is not best method to identify variance from middle value.
5. Normal distribution if not followed, model cannot be better.
6. Data distribution can be studied through **histogram or distribution plot** in seaborn.
7. Pandas pd to datetime, or datetime library
8. Dayfirst format, day,month year; default is month, day, year
9. Datetime conversion and extracting hour, time , seconds etc.( dt.hour,dt.time,dt.second)

|  |
| --- |
|  |
| Lecture 14 -10th October.  **Inferential Statistics**  Apply test / analysis on sample and not directly on population. And based on results apply on population.  Descriptive statistics: To understand more about sample collected. Sample is analysed using measure of central tendency, dispersion, distribution, skewness etc.  Inferential Statistices To make an estimate/ statement on the population with help of sample.  Bruke Force method: It is not possible to compute values on whole of population, so we take help of a sample .  Divisions of inferential statistics is Confidence value, Hypothesis testing and Statistical Testing  Steps involve  Getting a piece of data, analyse data, apply test, get results and apply on whole population  Central Limit Theorom: |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |

**07th November,2021**

Life cycle of datascientist, project flow, python programming functions, datastructures. Statistics, descriptive statistics, skewness, kurtosis, inferential statistics, hypothect. Machine learining about early predictions. Supervised, classification and regression. Output is available and parmetric or non parametric models. Parametric , simple linear regression, multiple linear regression

Evaluation matrix. AIC, BIC value.

Pvalue, whether it will contribute to prediction pvalue < 0.5

2)Accuracy score = True prediction/ overall prediction

==TP+TN/ TP+TN+FP+FN

3) Precision and recall of 1 relevant. To see stability of model prediction

TP/TP+FP Precision score close to 100 percentage.

4)Recall score speaks of individual class accuracy.

TP/TP+FN ( Ideal expectation =100 percentage)

Tradeoff between precision and recall. Should be close to 1.

Fi score to know harmonic mean of Precision and recall

Fiscore= 2\* Precision\*recall/ Precision +recall

ROC Curve plot between tp and fp

Roc – Receiver operating characteristics

plot between tp and fp

Auc=50%

Auc= Area under the curve needs to be close to 1. To know separability.

**Classification alogorithm**

**Binary classification or multiclass classification**

**Logistic regression**

**Sigmoid curve =** 1/1+e to power of -y where y=mx+e

Limit it between 0 and 1

Linear regression is the exact transformation of a linear regression.

Logistic regression gives more accurate value. is parametric and give slope and intercept.

Observation

Linear regression range is the probablity ranges between minus infinity to plus infinity

Logistic regression 0 to 1 with probability threshold 0.5 less than that no. more than that Yes.

Helps make yes or no predictin.

Confusion matrix to understand misclassification counts done by any model

Decision tree

Non parametric has no slope / interecept

All this for declaration of best model

**Kaggle:**  Coding for datascience.

Platform for practice datascience community

Companies post real business problem & datascience competition

Provides datasets of every possible field. Analyse past data so that prediction can be made

<https://www.kaggle.com->> Register-> Courses

Kaggle [somsocg@gmail.com](mailto:somsocg@gmail.com), Maiyu1293#

**x.shape,x\_train.shape,x\_test.shape**

((299, 12), (224, 12), (75, 12))

**y.shape,y\_train.shape,y\_test.shape**

((299, 1), (224, 1), (75, 1))

**fpr,tpr,thresholds=roc\_curve(y,logmodel.predict\_proba (x)[:1])**

**ValueError**: Found input variables with inconsistent numbers of samples: [299, 1]

RandomForest , Model

Generalized Model - Less Bias and Less Variance

Bagging, boosting, stacking.

Based on dataset size and variance algo to be decided.

Bagging( increases stability)

Sampling with replacement.

Row subsampling and Column subsampling.

Boosting algo(Gradient boosting)

increases accuracy.

Sequential classifier.

Stacking- havingobservations into c

Voting classifier

Regressors for continuos data.

Using different algorithms.,construct parallely.

Adaptive, gradient Light,Categorical boosting.

Extra tree classifier.

Sci-kit learn.org

Linear Regression

Logistic Regreesion’

Ada Boost, xg boost

Towards datascience

Analytical Vidya