Week 1

The first thing to do in any data science exercise is to **discover the problem**.

– 1) **Know the audience**(person/role) [bcz different audience means different questions on the same data]
– 2) Situation
– 3) Problem
– 4) Action
– 5) Impact

Eg.	John, the marketing head,	Person,role
	Must create a region-wise budget,	situation
	But doesn't know the region-wise ROI.	problem
	By prioritizing the region,	action
	He can maximize the ROI.	impact

Week 2

Get the data– by downloading, querying(using API), or scraping(from web pages or pdfs).

tsv– tab separated value

Libraries–

BeautifulSoup library- for scraping data from webpage
-to parse html code
geocoder Api using Nominatim is used for extracting location information.
requests library - to get webpage
json- to convert api into json dictionary object
urlencode- to give structure to long url
pandas - to manipulate the data
wikipedia- to extract data from wikipedia

```
wk.search ()
wk.summary()
wk.summary(.., sentences= limit number)
wk.page()
```

tabula- to convert pdf table to csv file

Import data from web to excel:-

Go to Data->new query->from other source-> from web
Data can be refreshed by just refreshing the table

Week 3 - Prepare/transform the data

Important questions:-

Q1 How does one preview the data?
Q2 How does one transform the data?
Q3 How does one clean the data?

1. Excel, Pandas profiling- load & preview the data
2. Google sheets, Excel, Trifecta's Wrangler- Transform the data, deriving matrices from data, adding columns
3. Python image library(Pillow)- transform image data
4. Tabula- extract data(tables) from pdf, OpenRefine- correct missing data, spelling mistakes in structured data
5. Excel- image labeling

Data aggregation is done to provide data summaries which help in examining trends, making comparisons or revealing interesting insights

Data aggregation with excel:-

To remove blanks from rows, use find&select and go to special and select blanks.

Color scale is an excel feature to identify clusters.

Sparkline- helps to find trends in data.
(insert->line->create sparkline)

Data bars- give graphical illustration to a column of *numbers*.
(conditional formatting-> data bar)

=Trim() function- to remove extra spaces

Pandas profiling library(Refer Video https://www.youtube.com/watch?v=CDwZPie29QQ):-

1) Generates a html report
2) Provide useful insights about the data like data distribution, variable outliers, correlation between variables.
3) For outliers/IQR— click toggle details-> quantile statistics

Openrefine (used for clustering)-

To correct spelling mistake

-go to column drop down->facet->text facet->cluster

-Clustering methods/algorithms used:-

- Key Collision (default)

-Nearest Neighbour: method Uses levenshtein distance(edit distance)

Levenshtein distance- no. of edits required to make two string same

Image labeling in Excel using macro code:-

 developer tab->design mode, insert-> image activex control->select area range

=Proper() is used to make words as only the first letter capitalized for each word.

Week 4 & 5

Pycaret library helps in building end to end machine learning pipelines.

EXCEL

Correlation- answers about how often is there an effect

Regression- answers about how much of an effect

Outliers- answers about when does this fail

PYTHON

Classification- answers about Which group does a given data point belong

Forecasting- answers about prediction

Clustering- answers about grouping of things (multiple variables)

Rattle- non-programmatic R based data modeling tool

PyCaret- automatic data modeling tool (provides models to choose from)

Regression in excel → data analysis toolpak

 Input Y range: dependent variable

 Input X range: independent variable

Simple Linear regression- one dependent variable and one independent variable

Multiple Linear Regression- one dependent variable and multiple independent variables, Adjusted R square (convert to %) is the amount of variation in dependent variable that can be explained by independent variables.

Significance value <0.05 ⇒ Good model

If P-value< 0.05, the independent variable is significant.

(Dependent variable) **Y** = **β**0*intercept +**β**1*1st independent variable +**β**2*second independent variable +**β**3*........

Rolling avg n= taking average of n previous days

For categorical variables where order can be specified -Ordinal encoder is used to convert them into numerical
For categorical variables where order can't be specified - Label encoder is used.

Image classification using Keras:-
keras.sequential() is a fully connected layer, New layers can be created such that all the neurons are connected
keras.functional()- the user should specify which neuron of the layer has to connect to which neurons of the next layer.

One hot encoding:-

1) From sklearn.preprocessing import OneHotEncoder
2) pandas.get_dummies

Kmeans(n), fit(..., sample_weight=...) : (skewed/weighted) clustering.
n is the no. of clusters to be formed

WEEK 6 & 7
Design the output– communicating the message to the audience through visuals
General Purpose tool– Excel, Google Data Studio, Power BI, Tableau
Specialized purpose tool– Excel(VBA), Flourish Studio(for better animation), Kumu(network visualization), QGIS(geographic visualization)

Q. How do you pick the right tool?
Ans. whatever you are most familiar with and that can get the job done, choose that tool.

Azure machine learning - Excel add-in
(column header for our data should be tweet_text)
Sentiment analysis helps us extract subjective information from text.
View schema provides info about type and format of input and output required
Predict- you specify the input and output range
Excel Azure ML add-in gives two outputs:- Sentiment and Score

Confusion matrix- provides Precision,Recall and Accuracy
Let TP= true positive , FN= False negative and so on…

Precision= Summation(TP) / Summation(TP + FP)
Recall= Summation(TP) / Summation(TP + FN)
Accuracy= Summation(TP + TN) / Summation(TP + FP + TN + FN)

Classification report- text blob sentiment prediction with respect to human labels
The Two apply functions:-
1)TextBlob_subjectivity score lies between 0 and 1
2)TextBlob_polarity score lies between -1 and 1. Around 0 means neutral.

Geopy computes **haversine** distance, geopy.distance.distance(..., …).km
Python library folium ((folium.Map) takes two parameters - location and zoom_start) is used to visualize geospatial data.

Flourish
Line chart duration(movement from one datapoint to another datapoint) is in milliseconds.
Line bar/pie template animates in terms of both drawing and morphing.
Survey- animation duration and stagger settings(in milliseconds)
Heatmap- fade and flip animation settings(in seconds)
Spider- animation duration for both 'draw' and 'morph'(in seconds)
Hierarchy- animation duration-simple speed settings(in seconds)

Google Charts

1. Extensive library of plots / charts
    a. Useful to browse around for inspiration on how to tell a story using your data (e.g., Sankey / alluvial charts)
2. User supplies information; Google charts returns graphical charts
3. Easy interface via R (googleVis)

Google Data Studio

1. Create simple dashboards / reports using basic interactive charts
2. Easier to learn compared to Tableau / Power Bi
3. … is free

Ideal for

1. Simple reports that require very little to no data processing / cleaning
2. High level reports that do not dive deep into finer details
3. When budget is a constraint

Four major concepts in google data studio:–
Connectors, Data Sources, Reports & Explorer
Data sources-Reports have many to many relationships
Connector is an interface of receiving data from various platforms.
Data source is a blueprint of data that can be modified
Explorer- to visualize data quickly

Tableau: one of the widely used enterprise software for data visualization.
Tableau Prep: etl tool for data engineering
Tableau Desktop: development tool for building interactive dashboards.
Tableau server: a hosting server to host large real time dashboards.

Kumu is a tool to organize complex data into relationship maps

Sankey diagram:- to visualize flow from one value to another.
Share google studio report: use owner's credentials
To get google charts source html code:- click Code it yourself JSFIDDLE

File format to create layers in QGIS= .shp (shape files)
Overlay shape file on top of world map:- use QuickMapServices plugin in QGIS application

Week 8
Tools to narrate story:- Excel, PowerBi, Tableau, Google sheets & Comicgen
Quill is used for narrating stories. It's an extension used in tableau. It transforms visualizations into narratives.
PowerBi :-

1) updates narrative automatically when data selection is modified.
2) allows custom narratives with dynamic values
3) use option "Summarise" to generate automatic narratives.

To enable Louvain algo:- from sknetwork.clustering import Louvain

Week 10

The final step in any data science process is deploying the results.
This has three parts (ML ops):-

- **Anonymize data**
  - To anonymize we use:
    - Tools (ARX, Amnesia)
    - or
    - Libraries (Faker, Mimesis)
- **Build app**
  - Notebook (Colab, Kaggle)
  - Data app (Streamlit(for python), Shiny(for R))
  - Web app (Flask, Tornado)
- **Host app**
  - Content (Github, Dropbox)
  - Apps (Heroku, Glitch)
  - Infra (AWS, Azure)

(Dev Ops)

- Secure app
- Scale app

Streamlit is a python library that allows us to deploy ML models by creating interfaces for displaying output.
-**ngrok** is a websock link that allows us to view/(expose) the link/(web server) that is created/(running) through colab in the local browser/ (in the local browser)

Heroku is a platform which allows us to publish web applications throughout the web.

Different Types of Hosting:-

Heroku

- Platform as a service
- Agile deployment for Ruby, Node.js, Clojure, Java, Python, Go and Scala.
- Run and scale any type of app.
- Total visibility across the entire app.
- Github integration.
- Command line interface.

- Recommended for ready-to-go machine learning model deployment for real time model inference and visualizations

Glitch

- User friendly browser text editor
- Difficulty in scaling.
- Low cache memory limit in runtime
- Github integration

Netlify

- Used for Static Web Hosting
- Global Network
- Instant Cache Validation
- Continuous deployments
- html/css/js backend
- Github integration
- Recommended for static websites

Vercel

- Used for Static Web Hosting
- Low cache memory limit in runtime
- Continuous deployments
- Github integration