

Assignment 2

Q.1 a) Explain how Watson Studio will provide the environment to solve business problems.

- ① Watson Studio provides you with the environment and tools to solve your business problems by collaboratively working with data.
- ② You can choose the tools you need to analyze and visualize data, to clean and shape data, to ingest streaming data, or to create and train machine learning models.
- ③ Visualizing information in graphical way can give you insights into your data.
- ④ By enabling you to look at and explore data from different perspectives, visualization can help you identify patterns, connections, and relationships within that data as well as understand large amounts of information very quickly.
- ⑤ Watson Studio is an IBM cloud-based platform designed to help businesses and data professionals through data science and machine learning.
- ⑥ It provides an environment that facilitates problem solving in several ways like data preparation, model development, scalability and etc.



Q.1 b) What is data types? List out the types of data types with example.

→ Data types describe the characteristic of a variable. It defines what type of data we are going to store in variable.

① Number: →

There are mainly three types which includes,

$a = 5$  → integer number.

$b = 2.5$  → float number.

$c = 6+2j$  → complex number.

② String: →

String is an ordered sequence of character. we can use single or double quotes to represent string. Example:  $str = \text{"welcome"}$

③ List: →

A list can contain a series of values. It is declared using `[]`. List is mutable, which means we can modify the list.

Example:  $list = [2, 3, 4, 5.5, 'Hi']$

④ Tuple: →

A tuple is a sequence of python objects separated by commas. Tuples are immutable, which



means tuples once created cannot be modified.

They are defined using parenthesis `()`.

Example: `tuple = (50, 25.6, "python")`

### ⑤ Set: →

A set is an unordered collection of items.

Set is defined by values separated by a comma inside braces `{ }`.

Example: `set = {5, 1, 2.6, "python"}`

### ⑥ Dictionary: →

Dictionary items are stored and fetched by using key. They are used to store huge amount of data. It contains key: value pairs.

Example: `dict = {1: "Hi", 2: 7.5, 3: "Hello"}`

Q.2a) What is dictionary? Explain the methods available in dictionary.

→ A) Dictionary are used to store data values in key-value pairs.

③ A dictionary is a collection which is ordered, changeable and does not allow duplicates.

### ⑤ Creating dictionary: →

With many key-value pairs surrounded in curly brackets and a colon separating each key.  
Syntax: `Dict = { "Name": "Sahil", "Roll No.": 140 }`



#### ④ Accessing the dictionary values : →

The key of dict can be used to obtain the values because they are unique from one another.

```
Ex. Emp = {"Name": "Dev", "Age": 20}
print("Name: %s" % Emp["Name"])
print("Age: %d" % Emp["Age"])
```

#### ⑤ Deleting element using pop() method is one of the way to get rid of elements from dict.

```
Ex. Dict = {1: "Hello", 2: "Key", 3: "Hi"}
pop-key = Dict.pop(2)
print(Dict)
```

#### ⑥ Iterating dict. can be iterated using 'for' loop

```
Ex. Emp = {"Name": "Sohil", "Age": 21}
for x in Emp:
    print(x)
```

#### ⑦ Delete keyword can be used to in-place delete key that is present in dict. in python.

```
dict = {"Arushi": 22, "Mansi": 21}
del dict["Mansi"]
```



Q. 2b) Differentiate bet<sup>n</sup> numpy arrays and lists.

→

Numpy Array

List

- ① It is the core library of python which is used for scientific computing. The core library of python provides list.
- ② It contains similar datatypes. It contains different datatypes.
- ③ It is homogeneous. It is both homogeneous as well as heterogeneous.
- ④ It is faster as compared to list and also has some optimisation function. It does not have optimisation function and also slow as compared to Numpy Array.
- ⑤ It requires smaller memory consumption as compared to python list. It requires more memory as compared to Numpy array.
- ⑥ In this element wise operation is available. Element wise list is not possible on the list.



Q. 26) Differentiate bet<sup>n</sup> numpy arrays and lists.

→

Numpy Array

List

① It is the core library of python which is used for scientific computing.

The core library of python provides list.

② It contain similar datatypes

It contain different datatypes

③ It is homogeneous

It is both homogeneous as well as heterogeneous.

④ It is faster as compared to list and also have some optimisation function

It does not have optimisation function and also slow as compared to Numpy Array.

⑤ It required smaller memory consumption as compared to python list

It required more memory as compared to Numpy array.

⑥ In this element wise operation is available

Element wise list is not possible on the list.



Q.3 a) Explain the following term:

i) Raw code : →

① In data visualisation, raw code typically refers to the direct programming code or scripts used to create visualizations from raw data.

② This code can be written in many programming languages like python, R and etc., depending on the visualization library or tool being used.

③ Raw code in data visualization provides flexibility and control over the design and presentation of visualisations.

ii) Waffle chart : →

① A waffle chart is an interesting visualization that is normally created to display progress towards goals.

② It is a commonly an effective option when you are trying to add "interesting" visualization features to a visual that consists mainly of cells, such as excel dashboard.



Q.3b) How will you create the normalized weight?

→ ① Identify the Range :- Determine the range of values in your dataset. For example, find minimum & maximum values.

② choose a Normalized method : There are various methods to normalize data, depending on your specific needs. Two common methods are:

a) min-max scaling :

$$\text{Normalized value} = \frac{\text{value} - \text{minimum}}{\text{maximum} - \text{minimum}}$$

b) Z-score stabilization :

$$\text{Normalized value} = \frac{\text{value} - \text{mean}}{\text{std. deviation}}$$

③ Apply the chosen method :- Use the chosen normalized method to scale your data. This will give you normalized values that represent the relative importance or weight of each data point within the chosen range.

④ Visualization : These normalized values will make it easier to compare and interpret the imp. of different data points or categories.



Q. 4 a) Explain pie chart specialized visualization tool using matplotlib.

→ ① Import matplotlib: Begin by importing matplotlib in your python script or Jupyter Notebook, import matplotlib.pyplot as plt.

② Prepare your data: you need to have the data you want to visualize in the form of numerical values. labels = ['category A', 'category B', 'category C', 'category D']  
sizes = [25, 30, 15, 30]

③ Create the piechart: use the plt.pie function to create the pie chart.

plt.pie(sizes, labels=labels, autopct='%1.1f%%',  
startangle=90, shadow=True)

'autopct': specifies the format for displaying the percentages on each slice.

'startangle': controls the starting angle for the first slice.

'shadow': Adds a shadow effect to the chart.

④ Display the pie chart: you can use the 'plt.show()' to display the piechart.  
plt.show()



Q.4 b) Explain Bubble plots specialized visualisation tool using matplotlib.

→ ① Import matplotlib: import matplotlib.pyplot as plt.

② prepare your data: you'll need data with three variables x-values, y-values and third variable that represent the size of the bubbles.

③ create the bubble plot: Example

$x = [1, 2, 3, 4, 5]$

$y = [10, 15, 13, 17, 20]$

$\text{bubble\_size} = [100, 200, 150, 250, 300]$

`plt.scatter(x, y, s=bubble_size, alpha=0.5, c='blue', label='Bubble plot')`

`plt.xlabel('x-axis label')`

`plt.ylabel('y-axis label')`

`plt.title("Bubble plot")`

`plt.legend`

`plt.show()`

④ In this example: "scatter" is used to create the bubble plot. The 's' parameter specified the size of each bubble based on the 'bubble-size' data. 'alpha' control the transparency of the bubbles. 'c' specifies the color of the bubbles. 'label' sets a label for legend.



Q.6 a) Explain the following term:

i) Regression plots: →

- The Regression plots in Seaborn are primarily intended to add a visual guide that help to emphasize pattern in a dataset during EDA.
- Regression plot as the name suggest creates a regression line between 2 parameters and help to visualise their linear relationship.
- Regression plots in Seaborn can be easily implemented with the help of `lmplot()` function

ii) Matrix plot: →

- Matrix plots, also known as heatmap matrices, are a type of data visualization techniques used to display the relationship bet<sup>n</sup> multiple variables in a dataset.
- They are particularly useful for exploring correlations or patterns in multivariate data.
- They can be used in various fields, including data analysis, biology and more to uncover hidden patterns.



Q.6 (b) Explain the following terms.

i) Distribution plots: →

- Distribution plot visually assesses the distribution of sample data by comparing the empirical distribution of the data with the theoretical values expected from a specified distribution.
- They are graphical representations that help you understand the distribution or spread of a dataset.

ii) Categorical plots: →

- Categorical plots are a type of data visualization used to display the distribution of data across different categories or groups.
- We have two different kinds of categorical distribution plots, box plot and violin plots.
- These kinds of plots allow us to choose a numerical variable, like age, and plot the distribution of age for each category in selected categorical variables.



Q.5 b) Explain spatial visualizations and analysis in python with folium.

→ ① Folium is a powerful data visualisation library in python that was built primarily to help people visualize geospatial data.

② With folium, you can create a map of any location in the world if you know its latitude and longitude value.

③ You can also create a map and superimpose markers as well as clusters of markers on top of the map for real and very interesting visualizations.

④ you can also create maps of different styles such as street level map, terrain map.

⑤ you can create a basic map by calling the function `folium.map()` and specifying the initial center and zoom level.

```
m = folium.map(location = [latitude, longitude],  
zoom_start = zoom_level)
```



Q.5a) An e-commerce company wants to get into logistics "delivery". It wants to know the pattern for maximum pickup call from different areas of the city throughout the day. This will result in:

i) Build optimum no. of stations where its pickup calls from different areas delivery personnel will be located.

ii) Ensure pickup personnel reaches the pickup location at the earliest possible time.

→ Find the highest density of probable pickup location in the future.

→ ① pre-requisites: python, Jupyter Notebook, pandas.

② Dataset: dataset contains two separate data files - train-del.csv and test-del.csv.

③ Importing libraries like pandas and folium and drop unnecessary attributes and combine 2 different files as one dataframe.

④ maps are defined as folium.map object. we will need to add other object on top of this before rendering.



⑤ visualise the rider data using a class method called Heatmap().

⑥ There is high demand for cabs in areas marked by the heat map which is central Delhi most probably and other surrounding areas.

⑦ we can also animate our heat maps to dynamically change the data on timely basis based on a certain dimension of time.

⑧ Conclusion:

Throughout the city, pickups are more probable from central areas so better to set lot of pickup stops at these location. Therefore, by using maps we can highlight trends and uncover pattern and derive insights from the data.