Data science

"The best way to learn data science is to apply data science."

Activity I

Bochra CHEMAM

I. Simple & Generic datasets to get you started:

- **Data.gov**
- **Five Thirty Eight Datasets**
- **Amazon Web Services (AWS) datasets**
- **Google datasets**
- **KDNuggets**

❖ **Datasets for predictive modelling & machine learning:**

- UCI Machine Learning Repository
- Kaggle

II. Application:

Today's activity aims to learn how we can apply the theoretical basics of statistics using python libraries such as:

1. **Python's statistics**: is a built-in Python library for descriptive statistics. You can use it if your datasets are not too large or if you can't rely on importing other libraries.

2. **NumPy:** is a third-party library for numerical computing, optimized for working with single- and multi-dimensional arrays. Its primary type is the array type called ndarray. This library contains many routines for statistical analysis.

3. **SciPy:** is a third-party library for scientific computing based on NumPy. It offers additional functionality compared to NumPy, including scipy.stats for statistical analysis.

4. **Pandas :** is a third-party library for numerical computing based on NumPy. It excels in handling labeled one-dimensional (1D) data with Series objects and two-dimensional (2D) data with DataFrame objects.

5. **Matplotlib :** is a third-party library for data visualization. It works well in combination with NumPy, SciPy, and Pandas.

6. **Seaborn**

Data science

MOHAMMED VI
POLYTECHNIC
UNIVERSITY

"The best way to learn data science is to apply data science."

We will **understand Descriptive Statistics using python :**

1- Brief introduction for statistics approach with a simple task:

- Create a random variable **X** with integer values between **0 and 20 for 50** observations.
- You can use **random.randint from numpy**
- Let's calculate Mean, Mode, and median and interpret results considering that :
  - ✓ **X: Results of exam for a class of 50 students**
- **Interpret results**

2- Download the data: we will use the dataset of Facebook users

(you can find it on Kaggle as well)

- Download your dataset using the **Pandas** library

- **Like** - The like which the user did.
- **LikesReceived** - Likes received by the user
- **Mobile-Likes** - Likes which user did on mobile
- **Mobile-LikesReceived** - Likes which user receive on mobile.
- **D.o.b** Date of Birthday
- **Tenure** - The number of days they have used Facebook (or spent on FB)

3- Data Structure Understand your data :

- Let's take a look at the top five rows using the DataFrames **head()**
- The **info()** method is useful to get a quick description of the data, in particular, the total number of rows, and each attribute's type, and the number of non-null values

- .isna().any() is useful to know if any nan value or we can use also df.isna().sum()/len(df)

4- Descriptive Statistics :

- **Use value_counts() method to know the different categories**
- **The describe() method shows a summary of the numerical attributes**

- The std row shows the standard deviation, which measures how dispersed the values are. The 25%, 50%, and 75% rows show the corresponding percentiles: a percentile indicates the value below which a given percentage of observations in a group of observations fall.
- For example, **25%** of the Facebook users are at or under 20 yrs while **50%** are lower than 28 and **75% are** lower than 50.
- These are often called the 25th percentile (or first quartile), the median, and the 75th percentile (or third quartile).

- Let's plot the age distribution of Facebook user using **Matplotlib** (Histogram)
- Let's plot Gender distribution using the same library but **with a pie chart**

- **Men tend to be on Facebook more often than women.**

- Let's plot tenure distribution using the same library histogram

- **the majority of the users were fairly new.**
  - You can use DF.hist()

5- **What do you notice from the plot?**

6- **Let plot Box plot using seaborn or matplotlib (gender by tenure/ gender by age …)**

- Use **groupby** to understand more the relation between columns

-Women have more friends on Facebook, **DF.groupby("gender")["friend_count"].mean()**

-they are also more likely to initiate friendship requests as well.

```
DF.groupby("gender")[ "friendships_initiated"].mean()
```

- **Finally, let's Plot the probability density function of tenure using seaborn**

  sns.kdeplot help us to plot the PDF of a continuous Variable