WORKING WITH THE DATASET

Sprint 01 : PREPARE THE DATASETS

DATE	10 NOV 2022
TEAM ID	PNT2022TMID00882
PROJECT NAME	Corporate Employee Attrition Analytics

PREPARE THE DATASET

```
[1] #Import Libraries
    import numpy as np
    import pandas as pd
    import seaborn as sns
```

```
[2] #Load the data
    from google.colab import files # Use to load data on Google Colab
    uploaded = files.upload() # Use to load data on Google Colab
```

```
Choose Files  IBM_HR_Tr...014-17.csv
• IBM_HR_Training_2014-17.csv(text/csv) - 106679 bytes, last modified: 10/14/2022 - 100% done
Saving IBM_HR_Training_2014-17.csv to IBM_HR_Training_2014-17.csv
```

```
[4] #Store the data into the df variable
    df = pd.read_csv('IBM_HR_Training_2014-17.csv')
    df.head(7) #Print the first 7 rows
```

	Year	Organization	Department	Position	Position count	Planned position count	Expense total	Course cost	Course days	Terminations	Internal hires	External hires
0	2014	GO Americas corporate	Human Resources	Vice-President of Human Resources	1.0	1.0	203,072.40	10,500	4.0	0.0	0.0	0.0
1	2014	GO Americas corporate	Human Resources	Human Resources Administrator	1.0	1.0	51,582.74	NaN	NaN	0.0	0.0	0.0
2	2014	GO Americas corporate	Human Resources	Benefit Specialist	2.0	2.0	79,115.38	NaN	NaN	0.0	0.0	0.0
3	2014	GO Americas corporate	Human Resources	Human Resources Clerk	2.0	2.0	65,658.44	NaN	NaN	0.0	0.0	0.0
4	2014	GO Americas corporate	Finance	Controller	1.0	1.0	205,051.67	10,500	4.0	0.0	0.0	0.0
5	2014	GO Americas corporate	Finance	Finance Manager	1.0	1.0	151,033.62	NaN	NaN	0.0	0.0	0.0
6	2014	GO Americas corporate	Finance	Financial Analyst	1.0	1.0	72,348.87	NaN	NaN	0.0	0.0	0.0

```
[5] #Get the number of rows and number of columns in the data
    df.shape
```

```
(1056, 12)
```

```
[6] #Count the empty (NaN, NAN, na) values in each column
    df.isna().sum()
```

```
Year                    0
Organization            0
Department              0
Position                0
Position count         24
Planned position count 24
Expense total          56
Course cost           431
Course days           431
Terminations           24
Internal hires         24
External hires         24
dtype: int64
```

```
df.isnull().values.any()
```

```
True
```